Econometrics

Badi H. Baltagi

Econometrics

Fourth Edition

Springer

Professor Badi H. Baltagi
Syracuse University
Center for Policy Research
426 Eggers Hall
Syracuse, NY 13244-1020
USA
bbaltagi@maxwell.syr.edu

ISBN 978-3-540-76515-8 e-ISBN 978-3-540-76516-5

DOI 10.1007/978-3-540-76516-5

Library of Congress Control Number: 2007939803

Production: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig
Coverdesign: WMX Design GmbH, Heidelberg

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

To My Wife Phyllis

Preface

This book is intended for a first year graduate course in econometrics. Courses requiring matrix algebra as a pre-requisite to econometrics can start with Chapter 7. Chapter 2 has a quick refresher on some of the required background needed from statistics for the proper understanding of the material in this book. For an advanced undergraduate/masters class not requiring matrix algebra, one can structure a course based on Chapter 1; Section 2.6 on descriptive statistics; Chapters 3-6; Section 11.1 on simultaneous equations; and Chapter 14 on time-series analysis.

This book teaches some of the basic econometric methods and the underlying assumptions behind them. Estimation, hypotheses testing and prediction are three recurrent themes in this book. Some uses of econometric methods include (i) empirical testing of economic theory, whether it is the permanent income consumption theory or purchasing power parity, (ii) forecasting, whether it is GNP or unemployment in the U.S. economy or future sales in the computer industry. (iii) Estimation of price elasticities of demand, or returns to scale in production. More importantly, econometric methods can be used to simulate the effect of policy changes like a tax increase on gasoline consumption, or a ban on advertising on cigarette consumption.

It is left to the reader to choose among the available econometric/statistical software to use, like EViews, SAS, STATA, TSP, SHAZAM, Microfit, PcGive, LIMDEP, and RATS, to mention a few. The empirical illustrations in the book utilize a variety of these software packages. Of course, these packages have different advantages and disadvantages. However, for the basic coverage in this book, these differences may be minor and more a matter of what software the reader is familiar or comfortable with. In most cases, I encourage my students to use more than one of these packages and to verify these results using simple programming languages like GAUSS, OX, R and MATLAB.

This book is not meant to be encyclopedic. I did not attempt the coverage of Bayesian econometrics simply because it is not my comparative advantage. The reader should consult Koop (2003) for a more recent treatment of the subject. Nonparametrics and semiparametrics are popular methods in today's econometrics, yet they are not covered in this book to keep the technical difficulty at a low level. These are a must for a follow-up course in econometrics, see Li and Racine (2007). Also, for a more rigorous treatment of asymptotic theory, see White (1984). Despite these limitations, the topics covered in this book are basic and necessary in the training of every economist. In fact, it is but a 'stepping stone', a 'sample of the good stuff' the reader will find in this young, energetic and ever evolving field.

I hope you will share my enthusiasm and optimism in the importance of the tools you will learn when you are through reading this book. Hopefully, it will encourage you to consult the suggested readings on this subject that are referenced at the end of each chapter. In his inaugural lecture at the University of Birmingham, entitled "Econometrics: A View from the Toolroom," Peter C.B. Phillips (1977) concluded:

> *"the toolroom may lack the glamour of economics as a practical art in government or business, but it is every bit as important. For the tools (econometricians) fashion provide the key to improvements in our quantitative information concerning matters of economic policy."*

As a student of econometrics, I have benefited from reading Johnston (1984), Kmenta (1986), Theil (1971), Klein (1974), Maddala (1977), and Judge, et al. (1985), to mention a few. As a teacher of undergraduate econometrics, I have learned from Kelejian and Oates (1989), Wallace and Silver (1988), Maddala (1992), Kennedy (1992), Wooldridge (2003) and Stock and Watson (2003). As a teacher of graduate econometrics courses, Greene (1993), Judge, et al. (1985), Fomby, Hill and Johnson (1984) and Davidson and MacKinnon (1993) have been my regular companions. The influence of these books will be evident in the pages that follow. At the end of each chapter I direct the reader to some of the classic references as well as further suggested readings.

This book strikes a balance between a rigorous approach that proves theorems and a completely empirical approach where no theorems are proved. Some of the strengths of this book lie in presenting some difficult material in a simple, yet rigorous manner. For example, Chapter 12 on pooling time-series of cross-section data is drawn from the author's area of expertise in econometrics and the intent here is to make this material more accessible to the general readership of econometrics.

The exercises contain theoretical problems that should supplement the understanding of the material in each chapter. Some of these exercises are drawn from the Problems and Solutions series of *Econometric Theory* (reprinted with permission of Cambridge University Press). In addition, the book has a set of empirical illustrations demonstrating some of the basic results learned in each chapter. Data sets from published articles are provided for the empirical exercises. These exercises are solved using several econometric software packages and are available in the Solution Manual. This book is by no means an applied econometrics text, and the reader should consult Berndt's (1991) textbook for an excellent treatment of this subject. Instructors and students are encouraged to get other data sets from the internet or journals that provide backup data sets to published articles. The *Journal of Applied Econometrics* and the *Journal of Business and Economic Statistics* are two such journals. In fact, the *Journal of Applied Econometrics* has a replication section for which I am serving as an editor. In my econometrics course, I require my students to replicate an empirical paper. Many students find this experience rewarding in terms of giving them hands on application of econometric methods that prepare them for doing their own empirical work.

I would like to thank my teachers Lawrence R. Klein, Roberto S. Mariano and Robert Shiller who introduced me to this field; James M. Griffin who provided some data sets, empirical exercises and helpful comments, and many colleagues who had direct and indirect influence on the contents of this book including G.S. Maddala, Jan Kmenta, Peter Schmidt, Cheng Hsiao, Tom Wansbeek, Walter Krämer, Maxwell King, Peter C.B. Phillips, Alberto Holly, Essie Maasoumi, Aris Spanos, Farshid Vahid, Heather Anderson, Arnold Zellner and Bryan Brown. Also, I would like to thank my students Wei-Wen Xiong, Ming-Jang Weng, Kiseok Nam, Dong Li and Gustavo Sanchez who read parts of this book and solved several of the exercises. Werner Müller and Martina Bihn at Springer for their prompt and professional editorial help. I have also benefited from my visits to the University of Arizona, University of California San-Diego, Monash University, the University of Zurich, the Institute of Advanced Studies in Vienna, and the University of Dortmund, Germany. A special thanks to my wife Phyllis whose help and support were essential to completing this book.

References

Berndt, E.R. (1991), *The Practice of Econometrics: Classic and Contemporary* (Addison-Wesley: Reading, MA).

Davidson, R. and J.G. MacKinnon (1993), *Estimation and Inference In Econometrics* (Oxford University Press: Oxford, MA).

Fomby, T.B., R.C. Hill and S.R. Johnson (1984), *Advanced Econometric Methods* (Springer-Verlag: New York).

Greene, W.H. (1993), *Econometric Analysis* (Macmillan: New York).

Johnston, J. (1984), *Econometric Methods* , 3rd. Ed., (McGraw-Hill: New York).

Judge, G.G., W.E. Griffiths, R.C. Hill, H. Lütkepohl and T.C. Lee (1985), *The Theory and Practice of Econometrics* , 2nd Ed., (John Wiley: New York).

Kelejian, H. and W. Oates (1989), *Introduction to Econometrics: Principles and Applications* , 2nd Ed., (Harper and Row: New York).

Kennedy, P. (1992), *A Guide to Econometrics* (The MIT Press: Cambridge, MA).

Klein, L.R. (1974), *A Textbook of Econometrics* (Prentice-Hall: New Jersey).

Kmenta, J. (1986), *Elements of Econometrics* , 2nd Ed., (Macmillan: New York).

Koop, G. (2003), *Bayesian Econometrics,* (Wiley: New York).

Li, Q. and J.S. Racine (2007), *Nonparametric Econometrics,* (Princeton University Press: New Jersey).

Maddala, G.S. (1977), *Econometrics* (McGraw-Hill: New York).

Maddala, G.S. (1992), *Introduction to Econometrics* (Macmillan: New York).

Phillips, P.C.B. (1977), "Econometrics: A View From the Toolroom," Inaugural Lecture, University of Birmingham, Birmingham, England.

Stock, J.H. and M.W. Watson (2003), *Introduction to Econometrics* , (Addison-Wesley: New York).

Theil, H. (1971), *Principles of Econometrics* (John Wiley: New York).

Wallace, T.D. and L. Silver (1988), *Econometrics: An Introduction* (Addison-Wesley: New York).

White, H. (1984), *Asymptotic Theory for Econometrics* (Academic Press: Florida).

Wooldridge, J.M. (2003), *Introductory Econometrics* , (South-Western: Ohio).

Data

The data sets used in this text can be downloaded from the Springer website in Germany. The address is: *http://www.springer.com/978-3-540-76515-8.* Please select the link "Samples & Supplements" from the right-hand column.

Table of Contents

Part I

CHAPTER 1
What Is Econometrics?

1.1 Introduction

What is econometrics? A few definitions are given below:

> *The method of econometric research aims, essentially, at a conjunction of economic theory and actual measurements, using the theory and technique of statistical inference as a bridge pier.*
>
> Trygve Haavelmo (1944)

> *Econometrics may be defined as the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference.*
>
> Samuelson, Koopmans and Stone (1954)

> *Econometrics is concerned with the systematic study of economic phenomena using observed data.*
>
> Aris Spanos (1986)

> *Broadly speaking, econometrics aims to give empirical content to economic relations for testing economic theories, forecasting, decision making, and for ex post decision/policy evaluation.*
>
> J. Geweke, J. Horowitz, and M.H. Pesaran (2007)

For other definitions of econometrics, see Tintner (1953).

An econometrician has to be a competent mathematician and statistician who is an economist by training. Fundamental knowledge of mathematics, statistics and economic theory are a necessary prerequisite for this field. As Ragnar Frisch (1933) explains in the first issue of Econometrica, it is the unification of statistics, economic theory and mathematics that constitutes econometrics. Each view point, by itself is necessary but not sufficient for a real understanding of quantitative relations in modern economic life.

Ragnar Frisch is credited with coining the term 'econometrics' and he is one of the founders of the Econometrics Society, see Christ (1983). Econometrics aims at giving empirical content to economic relationships. The three key ingredients are economic theory, economic data, and statistical methods. Neither 'theory without measurement', nor 'measurement without theory' are sufficient for explaining economic phenomena. It is as Frisch emphasized their union that is the key for success in the future development of econometrics.

Lawrence R. Klein, the 1980 recipient of the Nobel Prize in economics "for the creation of econometric models and their application to the analysis of economic fluctuations and economic policies,"[1] has always emphasized the integration of economic theory, statistical methods and practical economics. The exciting thing about econometrics is its concern for verifying or refuting economic laws, such as purchasing power parity, the life cycle hypothesis, the quantity theory of money, etc. These economic laws or hypotheses are testable with economic data. In fact, David F. Hendry (1980) emphasized this function of econometrics:

> *The three golden rules of econometrics are test, test and test; that all three rules are broken regularly in empirical applications is fortunately easily remedied. Rigorously tested models, which adequately described the available data, encompassed previous findings and were derived from well based theories would enhance any claim to be scientific.*

Econometrics also provides quantitative estimates of price and income elasticities of demand, estimates of returns to scale in production, technical efficiency, the velocity of money, etc. It also provides predictions about future interest rates, unemployment, or GNP growth. Lawrence Klein (1971) emphasized this last function of econometrics:

> *Econometrics had its origin in the recognition of empirical regularities and the systematic attempt to generalize these regularities into "laws" of economics. In a broad sense, the use of such "laws" is to make predictions - - about what might have or what will come to pass. Econometrics should give a base for economic prediction beyond experience if it is to be useful. In this broad sense it may be called the science of economic prediction.*

Econometrics, while based on scientific principles, still retains a certain element of art. According to Malinvaud (1966), the art in econometrics is trying to find the right set of assumptions which are sufficiently specific, yet realistic to enable us to take the best possible advantage of the available data. Data in economics are not generated under ideal experimental conditions as in a physics laboratory. This data cannot be replicated and is most likely measured with error. In some cases, the available data are proxies for variables that are either not observed or cannot be measured. Many published empirical studies find that economic data may not have enough variation to discriminate between two competing economic theories. Manski (1995, p. 8) argues that

> *Social scientists and policymakers alike seem driven to draw sharp conclusions, even when these can be generated only by imposing much stronger assumptions than can be defended. We need to develop a greater tolerance for ambiguity. We must face up to the fact that we cannot answer all of the questions that we ask.*

To some, the "art" element in econometrics has left a number of distinguished economists doubtful of the power of econometrics to yield sharp predictions. In his presidential address to the American Economic Association, Wassily Leontief (1971, pp. 2-3) characterized econometrics work as:

> *an attempt to compensate for the glaring weakness of the data base available to us by the widest possible use of more and more sophisticated techniques. Alongside the mounting pile of elaborate theoretical models we see a fast growing stock of equally intricate statistical tools. These are intended to stretch to the limit the meager supply of facts.*

Most of the time the data collected are not ideal for the economic question at hand because they were posed to answer legal requirements or comply to regulatory agencies. Griliches (1986, p.1466) describes the situation as follows:

> *Econometricians have an ambivilant attitude towards economic data. At one level, the 'data' are the world that we want to explain, the basic facts that economists purport to elucidate. At the other level, they are the source of all our trouble. Their imperfections make our job difficult and often impossible... We tend to forget that these imperfections are what gives us our legitimacy in the first place... Given that it is the 'badness' of the data that provides us with our living, perhaps it is not all that surprising that we have shown little interest in improving it, in getting involved in the grubby task of designing and collecting original data sets of our own. Most of our work is on 'found' data, data that have been collected by somebody else, often for quite different purposes.*

Even though economists are increasingly getting involved in collecting their data and measuring variables more accurately and despite the increase in data sets and data storage and computational accuracy, some of the warnings given by Griliches (1986, p. 1468) are still valid today:

> *The encounters between econometricians and data are frustrating and ultimately unsatisfactory both because econometricians want too much from the data and hence tend to be dissappointed by the answers, and because the data are incomplete and imperfect. In part it is our fault, the appetite grows with eating. As we get larger samples, we keep adding variables and expanding our models, until on the margin, we come back to the same insignificance levels.*

1.2 A Brief History

For a brief review of the origins of econometrics before World War II and its development in the 1940-1970 period, see Klein (1971). Klein gives an interesting account of the pioneering works of Moore (1914) on economic cycles, Working (1927) on demand curves, Cobb and Douglas (1928) on the theory of production, Schultz (1938) on the theory and measurement of demand, and Tinbergen (1939) on business cycles. As Klein (1971, p. 415) adds:

> *The works of these men mark the beginnings of formal econometrics. Their analysis was systematic, based on the joint foundations of statistical and economic theory, and they were aiming at meaningful substantive goals - to measure demand elasticity, marginal productivity and the degree of macroeconomic stability.*

The story of the early progress in estimating economic relationships in the U.S. is given in Christ (1985). The modern era of econometrics, as we know it today, started in the 1940's. Klein (1971) attributes the formulation of the econometrics problem in terms of the theory of statistical inference to Haavelmo (1943, 1944) and Mann and Wald (1943). This work was extended later by T.C. Koopmans, J. Marschak, L. Hurwicz, T.W. Anderson and others at the Cowles Commission in the late 1940's and early 1950's, see Koopmans (1950). Klein (1971, p. 416) adds:

> *At this time econometrics and mathematical economics had to fight for academic recognition. In retrospect, it is evident that they were growing disciplines and becoming increasingly attractive to the new generation of economic students after World War II, but only a few of the largest and most advanced universities offered formal work in these subjects. The mathematization of economics was strongly resisted.*

This resistance is a thing of the past, with econometrics being an integral part of economics, taught and practiced worldwide. Econometrica, the official journal of the Econometric Society is one of the leading journals in economics, and today the Econometric Society boast a large membership worldwide. Today, it is hard to read any professional article in leading economics and econometrics journals without seeing mathematical equations. Students of economics and econometrics have to be proficient in mathematics to comprehend this research. In an Econometric Theory interview, professor J. D. Sargan of the London School of Economics looks back at his own career in econometrics and makes the following observations: "... econometric theorists have really got to be much more professional statistical theorists than they had to be when I started out in econometrics in 1948... Of course this means that the starting econometrician hoping to do a Ph.D. in this field is also finding it more difficult to digest the literature as a prerequisite for his own study, and perhaps we need to attract students of an increasing degree of mathematical and statistical sophistication into our field as time goes by," see Phillips (1985, pp. 134-135). This is also echoed by another giant in the field, professor T.W. Anderson of Stanford, who said in an Econometric Theory interview: "These days econometricians are very highly trained in mathematics and statistics; much more so than statisticians are trained in economics; and I think that there will be more cross-fertilization, more joint activity," see Phillips (1986, p. 280).

Research at the Cowles Commission was responsible for providing formal solutions to the problems of identification and estimation of the simultaneous equations model, see Christ (1985).[2] Two important monographs summarizing much of the work of the Cowles Commission at Chicago, are Koopmans and Marschak (1950) and Koopmans and Hood (1953).[3] The creation of large data banks of economic statistics, advances in computing, and the general acceptance of Keynesian theory, were responsible for a great flurry of activity in econometrics. Macroeconometric modelling started to flourish beyond the pioneering macro models of Klein (1950) and Klein and Goldberger (1955).

For the story of the founding of Econometrica and the Econometric Society, see Christ (1983). Suggested readings on the history of econometrics are Pesaran (1987), Epstein (1987) and Morgan (1990). In the conclusion of her book on The History of Econometric Ideas, Morgan (1990; p. 264) explains:

> *In the first half of the twentieth century, econometricians found themselves carrying out a wide range of tasks: from the precise mathematical formulation of economic theories to the development tasks needed to build an econometric model; from the application of statistical methods in data preperation to the measurement and testing of models. Of necessity, econometricians were deeply involved in the creative development of both mathematical economic theory and statistical theory and techniques. Between the 1920s and the 1940s, the tools of mathematics and statistics were indeed used in a productive and complementary union to forge the essential ideas of the econometric approach. But the changing nature of the econometric enterprise in the 1940s caused a return to the division of labour favoured in the late nineteenth century, with mathematical economists working on theory building and econometricians concerned with statistical work. By the 1950s the founding ideal of econometrics, the union of mathematical and statistical economics into a truly synthetic economics, had collapsed.*

In modern day usage, econometrics have become the application of statistical methods to economics, like biometrics and psychometrics. Although, the ideals of Frisch still live on in Econometrica and the Econometric Society, Maddala (1999) argues that: "In recent years the issues of Econometrica have had only a couple of papers in econometrics (statistical methods in economics) and the rest are all on game theory and mathematical economics. If you look at the list of fellows of the Econometric Society, you find one or two econometricians and the rest are game theorists and mathematical economists." This may be a little exagerated but it does summarize the rift between modern day econometrics and mathematical economics. For a recent world wide ranking of econometricians as well as academic institutions in the field of econometrics, see Baltagi (2007).

1.3 Critiques of Econometrics

Econometrics has its critics. Interestingly, John Maynard Keynes (1940, p. 156) had the following to say about Jan Tinbergen's (1939) pioneering work:

> *No one could be more frank, more painstaking, more free of subjective bias or parti pris than Professor Tinbergen. There is no one, therefore, so far as human qualities go, whom it would be safer to trust with black magic. That there is anyone I would trust with it at the present stage or that this brand of statistical alchemy is ripe to become a branch of science, I am not yet persuaded. But Newton, Boyle and Locke all played with alchemy. So let him continue.*[4]

In 1969, Jan Tinbergen shared the first Nobel Prize in economics with Ragnar Frisch.

Recent well cited critiques of econometrics include the Lucas (1976) critique which is based on the Rational Expectations Hypothesis (REH). As Pesaran (1990, p. 17) puts it:

> *The message of the REH for econometrics was clear. By postulating that economic agents form their expectations endogenously on the basis of the true model of the economy and a correct understanding of the processes generating exogenous variables of the model, including government policy, the REH raised serious doubts about the invariance of the structural parameters of the mainstream macroeconometric models in face of changes in government policy.*

Responses to this critique include Pesaran (1987). Other lively debates among econometricians include Ed Leamer's (1983) article entitled "Let's Take the Con Out of Econometrics," and the response by McAleer, Pagan and Volker (1985). Rather than leave the reader with criticisms of econometrics especially before we embark on the journey to learn the tools of the trade, we conclude this section with the following quote from Pesaran (1990, pp. 25-26):

> *There is no doubt that econometrics is subject to important limitations, which stem largely from the incompleteness of the economic theory and the non-experimental nature of economic data. But these limitations should not distract us from recognizing the fundamental role that econometrics has come to play in the development of economics as a scientific discipline. It may not be possible conclusively to reject economic theories by means of econometric methods, but it does not mean that*

nothing useful can be learned from attempts at testing particular formulations of a given theory against (possible) rival alternatives. Similarly, the fact that econometric modelling is inevitably subject to the problem of specification searches does not mean that the whole activity is pointless. Econometric models are important tools for forecasting and policy analysis, and it is unlikely that they will be discarded in the future. The challenge is to recognize their limitations and to work towards turning them into more reliable and effective tools. There seem to be no viable alternatives.

1.4 Looking Ahead

Econometrics have experienced phenomenal growth in the past 50 years. There are five volumes of the *Handbook of Econometrics* running to 3833 pages. Most of it dealing with post 1960's research. A lot of the recent growth reflects the rapid advances in computing technology. The broad availability of micro data bases is a major advance which facilitated the growth of panel data methods (see Chapter 12) and microeconometric methods especially on sample selection and discrete choice (see Chapter 13) and that also lead to the award of the Nobel Prize in Economics to James Heckman and Daniel McFadden in 2000. The explosion in research in time series econometrics which lead to the development of ARCH and GARCH and cointegration (see Chapter 14) which also lead to the award of the Nobel Prize in Economics to Clive Granger and Robert Engle in 2003. It is a different world than it was 30 years ago. The computing facilities changed dramatically. The increasing accessibility of cheap and powerful computing facilities are helping to make the latest econometric methods more readily available to applied researchers. Today, there is hardly a field in economics which has not been intensive in its use of econometrics in empirical work. Pagan (1987, p. 81) observed that the work of econometric theorists over the period 1966-1986 have become part of the process of economic investigation and the training of economists. Based on this criterion, he declares econometrics as an "outstanding success." He adds that:

The judging of achievement inevitably involves contrast and comparison. Over a period of twenty years this would be best done by interviewing a time-travelling economist displaced from 1966 to 1986. I came into econometrics just after the beginning of this period, so have some appreciation for what has occurred. But because I have seen the events gradually unfolding, the effects upon me are not as dramatic. Nevertheless, let me try to be a time-traveller and comment on the perceptions of a 1966'er landing in 1986. My first impression must be of the large number of people who have enough econometric and computer skills to formulate, estimate and simulate highly complex and non-linear models. Someone who could do the equivalent tasks in 1966 was well on the way to a Chair. My next impression would be of the widespread use and purchase of econometric services in the academic, government, and private sectors. Quantification is now the norm rather than the exception. A third impression, gleaned from a sounding of the job market, would be a persistent tendency towards an excess demand for well-trained econometricians. The economist in me would have to acknowledge that the market judges the products of the discipline as a success.

The challenge for the 21^{st} century is to narrow the gap between theory and practice. Many feel that this gap has been widening with theoretical research growing more and more abstract and highly mathematical without an application in sight or a motivation for practical use. Heckman (2001) argues that econometrics is useful only if it helps economists conduct and interpret empirical research on economic data. He warns that the gap between econometric theory and empirical practice has grown over the past two decades. Theoretical econometrics becoming more closely tied to mathematical statistics. Although he finds nothing wrong, and much potential value, in using methods and ideas from other fields to improve empirical work in economics, he does warn of the risks involved in uncritically adopting the methods and mind set of the statisticians:

> *Econometric methods uncritically adapted from statistics are not useful in many research activities pursued by economists. A theorem-proof format is poorly suited for analyzing economic data, which requires skills of synthesis, interpretation and empirical investigation. Command of statistical methods is only a part, and sometimes a very small part, of what is required to do first class empirical research.*

In an Econometric Theory interview with Jan Tinbergen, Magnus and Morgan (1987, p.117) describe Tinbergen as one of the founding fathers of econometrics, publishing in the field from 1927 until the early 1950s. They add: "Tinbergen's approach to economics has always been a practical one. This was highly appropriate for the new field of econometrics, and enabled him to make important contributions to conceptual and theoretical issues, but always in the context of a relevant economic problem." The founding fathers of econometrics have always had the practitioner in sight. This is a far cry from many theoretical econometricians who refrain from applied work.

The recent entry by Geweke, Horowitz, and Pesaran (2007) in the *The New Palgrave Dictionary* provides the following recommendations for the future:

> *Econometric theory and practice seek to provide information required for informed decision-making in public and private economic policy. This process is limited not only by the adequacy of econometrics, but also by the development of economic theory and the adequacy of data and other information. Effective progress, in the future as in the past, will come from simultaneous improvements in econometrics, economic theory, and data. Research that specifically addresses the effectiveness of the interface between any two of these three in improving policy — to say nothing of all of them — necessarily transcends traditional subdisciplinary boundaries within economics. But it is precisely these combinations that hold the greatest promise for the social contribution of academic* economics.

Notes

1. See the interview of Professor L.R. Klein by Mariano (1987). Econometric Theory publishes interviews with some of the giants in the field. These interviews offer a wonderful glimpse at the life and work of these giants.

2. Simultaneous equations model is an integral part of econometrics and is studied in Chapter 11.

3. Tjalling Koopmans was the joint recipient of the Nobel Prize in Economics in 1975. In addition to his work on the identification and estimation of simultaneous equations models, he received the Nobel Prize for his work in optimization and economic theory.

4. I encountered this attack by Keynes on Tinbergen in the inaugural lecture that Peter C.B. Phillips (1977) gave at the University of Birmingham entitled "Econometrics: A View From the Toolroom," and David F. Hendry's (1980) article entitled "Econometrics - Alchemy or Science?"

References

Baltagi, B.H. (2007), "Worldwide Econometrics Rankings: 1989-2005," *Econometric Theory*, 23: 952-1012.

Christ, C.F. (1983), "The Founding of the Econometric Society and Econometrica," *Econometrica*, 51: 3-6.

Christ, C.F. (1985), "Early Progress in Estimating Quantiative Economic Relations in America," *American Economic Review*, 12: 39-52.

Cobb, C.W. and P.H. Douglas (1928), "A Theory of Production," *American Economic Review*, Supplement 18: 139-165.

Epstein, R.J. (1987), *A History of Econometrics* (North-Holland: Amsterdam).

Frisch, R. (1933), "Editorial," *Econometrica*, 1: 1-14.

Geweke, J., J. Horowitz, and M. H.Pesaran (2007), "Econometrics: A Bird's Eye View," forthcoming in *The New Palgrave Dictionary,* Second Edition.

Griliches, Z. (1986), "Economic Data Issues," in Z. Griliches and M.D. Intriligator (eds), *Handbook of Econometrics Vol. III* (North Holland: Amsterdam).

Haavelmo, T. (1943), "The Statistical Implications of a System of Simultaneous Equations," *Econometrica*, 11: 1-12.

Haavelmo, T. (1944), "The Probability Approach in Econometrics," *Econometrica*, Supplement to Volume 12: 1-118.

Heckman, J.J. (2001), "Econometrics and Empirical Economics," *Journal of Econometrics*, 100: 3-5.

Hendry, D.F. (1980), "Econometrics - Alchemy or Science?" *Economica*, 47: 387-406.

Keynes, J.M. (1940), "On Method of Statistical Research: Comment," *Economic Journal*, 50: 154-156.

Klein, L.R. (1971), "Whither Econometrics?" *Journal of the American Statistical Association*, 66: 415-421.

Klein, L.R. (1950), *Economic Fluctuations in the United States 1921-1941*, Cowles Commission Monograph, No. 11 (John Wiley: New York).

Klein, L.R. and A.S. Goldberger (1955), *An Econometric Model of the United States 1929-1952* (North-Holland: Amsterdam).

Koopmans, T.C. (1950), ed., *Statistical Inference in Dynamic Economic Models* (John Wiley: New York).

Koopmans, T.C. and W.C. Hood (1953), *Studies in Econometric Method* (John Wiley: New York).

Koopmans, T.C. and J. Marschak (1950), eds., *Statistical Inference in Dynamic Economic Models* (John Wiley: New York).

Leamer, E.E. (1983), "Lets Take the Con Out of Econometrics," *American Economic Review*, 73: 31-43.

Leontief, W. (1971), "Theoretical Assumptions and Nonobserved Facts," *American Economic Review*, 61: 1-7.

Lucas, R.E. (1976), "Econometric Policy Evaluation: A Critique," in K. Brunner and A.M. Meltzer, eds., *The Phillips Curve and Labor Markets*, Carnegie Rochester Conferences on Public Policy, 1: 19-46.

Maddala, G.S. (1999), "Econometrics in the 21^{st} Century," in C.R. Rao and R. Szekeley, eds., *Statistics for the 21^{st} Century* (Marcel Dekker: New York).

Magnus , J.R. and M.S. Morgan (1987), "The ET Interview: Professor J. Tinbergen," *Econometric Theory*, 3: 117-142.

Malinvaud, E. (1966), *Statistical Methods of Econometrics* (North-Holland: Amsterdam).

Manski, C.F. (1995), *Identification Problems in the Social Sciences* (Harvard University Press: Cambridge).

Mann, H.B. and A. Wald (1943), "On the Statistical Treatment of Linear Stochastic Difference Equations," *Econometrica*, 11: 173-220.

Mariano, R.S. (1987), "The ET Interview: Professor L.R. Klein," *Econometric Theory*, 3: 409-460.

McAleer, M., A.R. Pagan and P.A. Volker (1985), "What Will Take The Con Out of Econometrics," *American Economic Review*, 75: 293-307.

Moore, H.L. (1914), *Economic Cycles: Their Law and Cause* (Macmillan: New York).

Morgan, M. (1990), *The History of Econometric Ideas* (Cambridge University Press: Cambridge, MA).

Pagan, A. (1987), "Twenty Years After: Econometrics, 1966-1986," paper presented at CORE's 20th Anniversary Conference, Louvain-la-Neuve.

Pesaran, M.H. (1987), *The Limits to Rational Expectations* (Basil Blackwell: Oxford, MA).

Pesaran, M.H. (1990), "Econometrics," in J. Eatwell, M. Milgate and P. Newman; *The New Palgrave: Econometrics* (W.W. Norton and Company: New York).

Phillips, P.C.B. (1977), "Econometrics: A View From the Toolroom," *Inaugural Lecture*, University of Birmingham, Birmingham, England.

Phillips, P.C.B. (1985), "ET Interviews: Professor J. D. Sargan," *Econometric Theory*, 1: 119-139.

Phillips, P.C.B. (1986), "The ET Interview: Professor T. W. Anderson," *Econometric Theory*, 2: 249-288.

Samuelson, P.A., T.C. Koopmans and J.R.N. Stone (1954), "Report of the Evaluative Committee for Econometrica," *Econometrica*, 22: 141-146.

Schultz, H. (1938), *The Theory and Measurement of Demand* (University of Chicago Press: Chicago, IL).

Spanos, A. (1986), *Statistical Foundations of Econometric Modelling* (Cambridge University Press: Cambridge, MA).

Tinbergen, J. (1939), *Statistical Testing of Business Cycle Theories, Vol. II: Business Cycles in the USA, 1919-1932* (League of Nations: Geneva).

Tintner, G. (1953), "The Definition of Econometrics," *Econometrica*, 21: 31-40.

Working, E.J. (1927), "What Do Statistical 'Demand Curves' Show?" *Quarterly Journal of Economics*, 41: 212-235.

CHAPTER 2
Basic Statistical Concepts

2.1 Introduction

One chapter cannot possibly review what one learned in one or two pre-requisite courses in statistics. This is an econometrics book, and it is imperative that the student have taken at least one solid course in statistics. The concepts of a random variable, whether discrete or continuous, and the associated probability function or *probability density function* (p.d.f.) are assumed known. Similarly, the reader should know the following statistical terms: Cumulative distribution function, marginal, conditional and joint p.d.f.'s. The reader should be comfortable with computing mathematical expectations, and familiar with the concepts of independence, Bayes Theorem and several continuous and discrete probability distributions. These distributions include: the Bernoulli, Binomial, Poisson, Geometric, Uniform, Normal, Gamma, Chi-squared (χ^2), Exponential, Beta, t and F distributions.

Section 2.2 reviews two methods of estimation, while section 2.3 reviews the properties of the resulting estimators. Section 2.4 gives a brief review of test of hypotheses, while section 2.5 discusses the meaning of confidence intervals. These sections are fundamental background for this book, and the reader should make sure that he or she is familiar with these concepts. Also, be sure to solve the exercises at the end of this chapter.

2.2 Methods of Estimation

Consider a Normal distribution with mean μ and variance σ^2. This is the important "Gaussian" distribution which is symmetric and bell-shaped and completely determined by its measure of centrality, its mean μ and its measure of dispersion, its variance σ^2. μ and σ^2 are called the population parameters. Draw a random sample $X_1, \ldots, X_n$ independent and identically distributed (IID) from this population. We usually estimate μ by $\widehat{\mu} = \bar{X}$ and σ^2 by

$$s^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1).$$

For example, μ = mean income of a household in Houston. $\bar{X}$ = sample average of incomes of 100 households randomly interviewed in Houston.

This estimator of μ could have been obtained by either of the following two methods of estimation:

(i) Method of Moments

Simply stated, this method of estimation uses the following rule: Keep equating population moments to their sample counterpart until you have estimated all the population parameters.

Population	Sample
$E(X) = \mu$	$\sum_{i=1}^{n} X_i/n = \bar{X}$
$E(X^2) = \mu^2 + \sigma^2$	$\sum_{i=1}^{n} X_i^2/n$
$\vdots$	$\vdots$
$E(X^r)$	$\sum_{i=1}^{n} X_i^r/n$

The normal density is completely identified by μ and σ^2, hence only the first 2 equations are needed

$$\widehat{\mu} = \bar{X} \qquad \text{and} \qquad \widehat{\mu}^2 + \widehat{\sigma}^2 = \sum_{i=1}^{n} X_i^2/n$$

Substituting the first equation in the second one obtains

$$\widehat{\sigma}^2 = \sum_{i=1}^{n} X_i^2/n - \bar{X}^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/n$$

(ii) Maximum Likelihood Estimation (MLE)

For a random sample of size n from the Normal distribution $X_i \sim N(\mu, \sigma^2)$, we have

$$f_i(X_i; \mu, \sigma^2) = (1/\sigma\sqrt{2\pi}) \exp\left\{-(X_i - \mu)^2/2\sigma^2\right\} \quad -\infty < X_i < +\infty$$

Since $X_1, \ldots, X_n$ are independent and identically distributed, the joint probability density function is given as the product of the marginal probability density functions:

$$f(X_1, \ldots, X_n; \mu, \sigma^2) = \prod_{i=1}^{n} f_i(X_i; \mu, \sigma^2) = (1/2\pi\sigma^2)^{n/2} \exp\left\{-\sum_{i=1}^{n} (X_i - \mu)^2/2\sigma^2\right\} \quad (2.1)$$

Usually, we observe only one sample of n households which could have been generated by any pair of (μ, σ^2) with $-\infty < \mu < +\infty$ and $\sigma^2 > 0$. For each pair, say (μ_0, σ_0^2), $f(X_1, \ldots, X_n; \mu_0, \sigma_0^2)$ denotes the probability (or likelihood) of obtaining that sample. By varying (μ, σ^2) we get different probabilities of obtaining this sample. Intuitively, we choose the values of μ and σ^2 that maximize the probability of obtaining this sample. Mathematically, we treat $f(X_1, \ldots, X_n; \mu, \sigma^2)$ as $L(\mu, \sigma^2)$ and we call it the likelihood function. Maximizing $L(\mu, \sigma^2)$ with respect to μ and σ^2, one gets the first-order conditions of maximization:

$$(\partial L/\partial \mu) = 0 \qquad \text{and} \qquad (\partial L/\partial \sigma^2) = 0$$

Equivalently, we can maximize $\log L(\mu, \sigma^2)$ rather than $L(\mu, \sigma^2)$ and still get the same answer. Usually, the latter monotonic transformation of the likelihood is easier to maximize and the first-order conditions become

$$(\partial \log L/\partial \mu) = 0 \qquad \text{and} \qquad (\partial \log L/\partial \sigma^2) = 0$$

For the Normal distribution example, we get

$$\log L(\mu; \sigma^2) = -(n/2)\log\, \sigma^2 - (n/2)\log\, 2\pi - (1/2\sigma^2) \sum_{i=1}^{n} (X_i - \mu)^2$$

$$\partial \log L(\mu; \sigma^2)/\partial \mu = (1/\sigma^2) \sum_{i=1}^{n} (X_i - \mu) = 0 \Rightarrow \widehat{\mu}_{MLE} = \bar{X}$$

$$\partial \log L(\mu; \sigma^2)/\partial \sigma^2 = -(n/2)(1/\sigma^2) + \sum_{i=1}^{n} (X_i - \mu)^2/2\sigma^4 = 0$$

$$\Rightarrow \widehat{\sigma}^2_{MLE} = \sum_{i=1}^{n} (X_i - \widehat{\mu}_{MLE})^2/n = \sum_{i=1}^{n} (X_i - \bar{X})^2/n$$

Note that the moments estimators and the maximum likelihood estimators are the same for the Normal distribution example. In general, the two methods need not necessarily give the same estimators. Also, note that the moments estimators will always have the same estimating equations, for example, the first two equations are always

$$E(X) = \mu \equiv \sum_{i=1}^{n} X_i/n = \bar{X} \qquad \text{and} \qquad E(X^2) = \mu^2 + \sigma^2 \equiv \sum_{i=1}^{n} X_i^2/n.$$

For a specific distribution, we need only substitute the relationship between the population moments and the parameters of that distribution. Again, the number of equations needed depends upon the number of parameters of the underlying distribution. For e.g., the exponential distribution has one parameter and needs only one equation whereas the gamma distribution has two parameters and needs two equations. Finally, note that the maximum likelihood technique is heavily reliant on the form of the underlying distribution, but it has desirable properties when it exists. These properties will be discussed in the next section.

So far we have dealt with the Normal distribution to illustrate the two methods of estimation. We now apply these methods to the Bernoulli distribution and leave other distributions applications to the exercises. We urge the student to practice on these exercises.

Bernoulli Example: In various cases in real life the outcome of an event is binary, a worker may join the labor force or may not. A criminal may return to crime after parole or may not. A television off the assembly line may be defective or not. A coin tossed comes up head or tail, and so on. In this case $\theta = \Pr[\text{Head}]$ and $1 - \theta = \Pr[\text{Tail}]$ with $0 < \theta < 1$ and this can be represented by the discrete probability function

$$\begin{array}{lll} f(X;\theta) & = \theta^X (1-\theta)^{1-X} & X = 0, 1 \\ & = 0 & \text{elsewhere} \end{array}$$

The Normal distribution is a continuous distribution since it takes values for all X over the real line. The Bernoulli distribution is discrete, because it is defined only at integer values for X. Note that $P[X = 1] = f(1;\theta) = \theta$ and $P[X = 0] = f(0;\theta) = 1 - \theta$ for all values of $0 < \theta < 1$. A random sample of size n drawn from this distribution will have a joint probability function

$$L(\theta) = f(X_1, \ldots, X_n; \theta) = \theta^{\sum_{i=1}^{n} X_i} (1-\theta)^{n - \sum_{i=1}^{n} X_i}$$

with $X_i = 0, 1$ for $i = 1, \ldots, n$. Therefore,

$$\begin{array}{rcl} \log L(\theta) & = & (\sum_{i=1}^{n} X_i) \log\theta + (n - \sum_{i=1}^{n} X_i) \log(1-\theta) \\ \dfrac{\partial \log L(\theta)}{\partial \theta} & = & \dfrac{\sum_{i=1}^{n} X_i}{\theta} - \dfrac{(n - \sum_{i=1}^{n} X_i)}{(1-\theta)} \end{array}$$

Solving this first-order condition for θ, one gets

$$(\sum_{i=1}^{n} X_i)(1-\theta) - \theta(n - \sum_{i=1}^{n} X_i) = 0$$

which reduces to

$$\widehat{\theta}_{MLE} = \sum_{i=1}^{n} X_i/n = \bar{X}.$$

This is the frequency of heads in n tosses of a coin.

For the method of moments, we need

$$E(X) = \sum_{X=0}^{1} Xf(X,\theta) = 1.f(1,\theta) + 0.f(0,\theta) = f(1,\theta) = \theta$$

and this is equated to $\bar{X}$ to get $\widehat{\theta} = \bar{X}$. Once again, the MLE and the method of moments yield the same estimator. Note that only one parameter θ characterizes this Bernoulli distribution and one does not need to equate second or higher population moments to their sample values.

2.3 Properties of Estimators

(i) Unbiasedness

$\widehat{\mu}$ is said to be unbiased for μ if and only if $E(\widehat{\mu}) = \mu$

For $\widehat{\mu} = \bar{X}$, we have $E(\bar{X}) = \sum_{i=1}^{n} E(X_i)/n = \mu$ and $\bar{X}$ is unbiased for μ. No distributional assumption is needed as long as the X_i's are distributed with the same mean μ. Unbiasedness means that "on the average" our estimator is on target. Let us explain this last statement. If we repeat our drawing of a random sample of 100 households, say 200 times, then we get 200 $\bar{X}$'s. Some of these $\bar{X}$'s will be above μ some below μ, but their average should be very close to μ. Since in real life situations, we observe only one random sample, there is little consolation if our observed $\bar{X}$ is far from μ. But the larger n is the smaller is the dispersion of this $\bar{X}$, since $\text{var}(\bar{X}) = \sigma^2/n$ and the lesser is the likelihood of this $\bar{X}$ to be very far from μ. This leads us to the concept of efficiency.

(ii) Efficiency

For two unbiased estimators, we compare their efficiencies by the ratio of their variances. We say that the one with lower variance is more efficient. For example, taking $\widehat{\mu}_1 = X_1$ versus $\widehat{\mu}_2 = \bar{X}$, both estimators are unbiased but $\text{var}(\widehat{\mu}_1) = \sigma^2$ whereas, $\text{var}(\widehat{\mu}_2) = \sigma^2/n$ and {the relative efficiency of $\widehat{\mu}_1$ with respect to $\widehat{\mu}_2$} $= \text{var}(\widehat{\mu}_2)/\text{var}(\widehat{\mu}_1) = 1/n$, see Figure 2.1. To compare all unbiased estimators, we find the one with minimum variance. Such an estimator if it exists is called the MVU (minimum variance unbiased estimator). A lower bound for the variance of any unbiased estimator $\widehat{\mu}$ of μ, is known in the statistical literature as the Cramér-Rao lower bound, and is given by

$$\text{var}(\widehat{\mu}) \geq 1/n\{E(\partial \log f(X;\mu))/\partial\mu\}^2 = -1/\{nE(\partial^2 \log f(X;\mu))/\partial\mu^2\} \tag{2.2}$$

where we use either representation of the bound on the right hand side of (2.2) depending on which one is the simplest to derive.

Example 1: Consider the normal density

$$\log f(X_i;\mu) = (-1/2)\log\sigma^2 - (1/2)\log 2\pi - (1/2)(X_i - \mu)^2/\sigma^2$$

$$\partial \log f(X_i;\mu)/\partial\mu = (X_i - \mu)/\sigma^2$$

$$\partial^2 \log f(X_i;\mu)/\partial\mu^2 = -(1/\sigma^2)$$

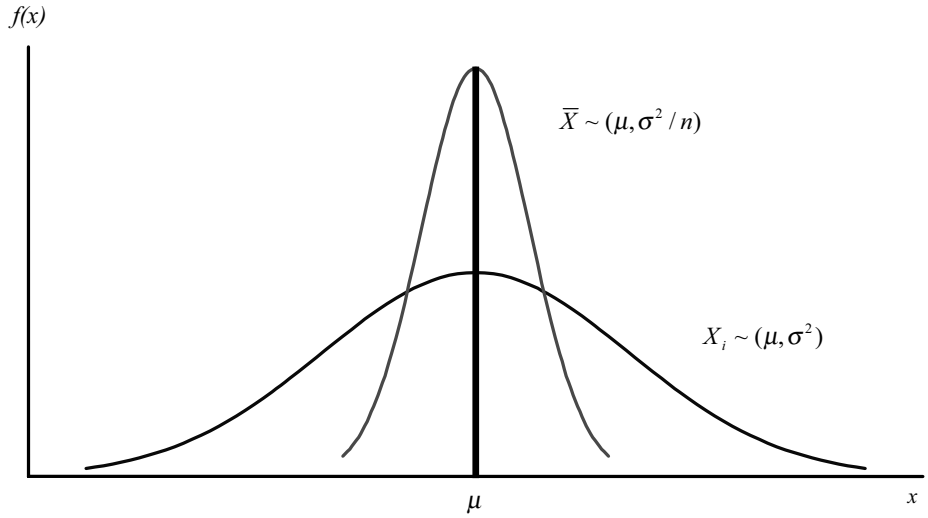


Figure 2.1 Efficiency Comparisons

with $E\{\partial^2 \log f(X_i;\mu)/\partial\mu^2\} = -(1/\sigma^2)$. Therefore, the variance of any unbiased estimator of μ, say $\widehat{\mu}$ satisfies the property that $\text{var}(\widehat{\mu}) \geq \sigma^2/n$.

Turning to σ^2; let $\theta = \sigma^2$, then

$$\log f(X_i;\theta) = -(1/2)\log\theta - (1/2)\log 2\pi - (1/2)(X_i-\mu)^2/\theta$$

$$\partial \log f(X_i;\theta)/\partial\theta = -1/2\theta + (X_i-\mu)^2/2\theta^2 = \{(X_i-\mu)^2 - \theta\}/2\theta^2$$

$$\partial^2 \log f(X_i;\theta)/\partial\theta^2 = 1/2\theta^2 - (X_i-\mu)^2/\theta^3 = \{\theta - 2(X_i-\mu)^2\}/2\theta^3$$

$E[\partial^2 \log f(X_i;\theta)/\partial\theta^2] = -(1/2\theta^2)$, since $\text{E}(X_i-\mu)^2 = \theta$. Hence, for any unbiased estimator of θ, say $\widehat{\theta}$, its variance satisfies the following property $\text{var}(\widehat{\theta}) \geq 2\theta^2/n$, or $\text{var}(\widehat{\sigma}^2) \geq 2\sigma^4/n$.

Note that, if one finds an unbiased estimator whose variance attains the Cramér-Rao lower bound, then this is the MVU estimator. It is important to remember that this is only a lower bound and sometimes it is not necessarily attained. If the X_i's are normal, $\bar{X} \sim N(\mu, \sigma^2/n)$. Hence, $\bar{X}$ is unbiased for μ with variance σ^2/n equal to the Cramér-Rao lower bound. Therefore, $\bar{X}$ is MVU for μ. On the other hand,

$$\widehat{\sigma}^2_{MLE} = \textstyle\sum_{i=1}^{n}(X_i - \bar{X})^2/n,$$

and it can be shown that $(n\widehat{\sigma}^2_{MLE})/(n-1) = s^2$ is unbiased for σ^2. In fact, $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ and the expected value of a Chi-squared variable with $(n-1)$ degrees of freedom is exactly its degrees of freedom. Using this fact,

$$E\{(n-1)s^2/\sigma^2\} = E(\chi^2_{n-1}) = n-1.$$

Therefore, $E(s^2) = \sigma^2$.[1] Also, the variance of a Chi-squared variable with $(n-1)$ degrees of freedom is twice these degrees of freedom. Using this fact,

$$\text{var}\{(n-1)s^2/\sigma^2\} = \text{var}(\chi^2_{n-1}) = 2(n-1)$$

or

$$\{(n-1)^2/\sigma^4\}\text{var}(s^2) = 2(n-1).$$

Hence, the $\text{var}(s^2) = 2\sigma^4/(n-1)$ and this does not attain the Cramér-Rao lower bound. In fact, it is larger than $(2\sigma^4/n)$. Note also that $\text{var}(\widehat{\sigma}^2_{MLE}) = \{(n-1)^2/n^2\}\text{var}(s^2) = \{2(n-1)\}\sigma^4/n^2$. This is smaller than $(2\sigma^4/n)$! How can that be? Remember that $\widehat{\sigma}^2_{MLE}$ is a biased estimator of σ^2 and hence, $\text{var}(\widehat{\sigma}^2_{MLE})$ should not be compared with the Cramér-Rao lower bound. This lower bound pertains only to unbiased estimators.

Warning: Attaining the Cramér-Rao lower bound is only a sufficient condition for efficiency. Failing to satisfy this condition does not necessarily imply that the estimator is not efficient.

Example 2: For the Bernoulli case

$$\log f(X_i;\theta) = X_i \log\theta + (1-X_i)\log(1-\theta)$$

$$\partial \log f(X_i,\theta)/\partial\theta = (X_i/\theta) - (1-X_i)/(1-\theta)$$

$$\partial^2 \log f(X_i;\theta)/\partial\theta^2 = (-X_i/\theta^2) - (1-X_i)/(1-\theta)^2$$

and $E[\partial^2 \log f(X_i;\theta)/\partial\theta^2] = (-1/\theta) - 1/(1-\theta) = -1/[\theta(1-\theta)]$. Therefore, for any unbiased estimator of θ, say $\widehat{\theta}$, its variance satisfies the following property:

$$\text{var}(\widehat{\theta}) \geq \theta(1-\theta)/n.$$

For the Bernoulli random sample, we proved that $\mu = E(X_i) = \theta$. Similarly, it can be easily verified that $\sigma^2 = \text{var}(X_i) = \theta(1-\theta)$. Hence, $\bar{X}$ has mean $\mu = \theta$ and $\text{var}(\bar{X}) = \sigma^2/n = \theta(1-\theta)/n$. This means that $\bar{X}$ is unbiased for θ and it attains the Cramér-Rao lower bound. Therefore, $\bar{X}$ is MVU for θ.

Unbiasedness and efficiency are finite sample properties (in other words, true for any finite sample size n). Once we let n tend to ∞ then we are in the realm of *asymptotic properties*.

Example 3: For a random sample from any distribution with mean μ it is clear that $\widetilde{\mu} = (\bar{X} + 1/n)$ is not an unbiased estimator of μ since $E(\widetilde{\mu}) = E(\bar{X} + 1/n) = \mu + 1/n$. However, as $n \to \infty$ the lim $E(\widetilde{\mu})$ is equal to μ. We say, that $\widetilde{\mu}$ is *asymptotically unbiased* for μ.

Example 4: For the Normal case

$$\widehat{\sigma}^2_{MLE} = (n-1)s^2/n \quad \text{and} \quad E(\widehat{\sigma}^2_{MLE}) = (n-1)\sigma^2/n.$$

But as $n \to \infty$, lim $E(\widehat{\sigma}^2_{MLE}) = \sigma^2$. Hence, $\widehat{\sigma}^2_{MLE}$ is *asymptotically unbiased* for σ^2.

Similarly, an estimator which attains the Cramér-Rao lower bound in the limit is *asymptotically efficient*. Note that $\text{var}(\bar{X}) = \sigma^2/n$, and this tends to zero as $n \to \infty$. Hence, we consider $\sqrt{n}\bar{X}$ which has finite variance since $\text{var}(\sqrt{n}\bar{X}) = n\ \text{var}(\bar{X}) = \sigma^2$. We say that the asymptotic variance of $\bar{X}$ denoted by $\text{asymp.var}(\bar{X}) = \sigma^2/n$ and that it attains the Cramér-Rao lower bound in the limit. $\bar{X}$ is therefore asymptotically efficient. Similarly,

$$\text{var}(\sqrt{n}\widehat{\sigma}^2_{MLE}) = n\ \text{var}(\widehat{\sigma}^2_{MLE}) = 2(n-1)\sigma^4/n$$

which tends to $2\sigma^4$ as $n \to \infty$. This means that $\text{asymp.var}(\widehat{\sigma}^2_{MLE}) = 2\sigma^4/n$ and that it attains the Cramér-Rao lower bound in the limit. Therefore, $\widehat{\sigma}^2_{MLE}$ is asymptotically efficient.

(iii) Consistency

Another asymptotic property is consistency. This says that as $n \to \infty$ $\lim \Pr[|\bar{X} - \mu| > c] = 0$ for any arbitrary positive constant c. In other words, $\bar{X}$ will not differ from μ as $n \to \infty$.

Proving this property uses the Chebyshev's inequality which states in this context that

$$\Pr[|\bar{X} - \mu| > k\sigma_{\bar{X}}] \leq 1/k^2.$$

If we let $c = k\sigma_{\bar{X}}$ then $1/k^2 = \sigma^2_{\bar{X}}/c^2 = \sigma^2/nc^2$ and this tends to 0 as $n \to \infty$, since σ^2 and c are finite positive constants. A sufficient condition for an estimator to be consistent is that it is asymptotically unbiased and that its variance tends to zero as $n \to \infty$.[2]

Example 1: For a random sample from *any* distribution with mean μ and variance σ^2, $E(\bar{X}) = \mu$ and $\text{var}(\bar{X}) = \sigma^2/n \to 0$ as $n \to \infty$, hence $\bar{X}$ is consistent for μ.

Example 2: For the Normal case, we have shown that $E(s^2) = \sigma^2$ and $\text{var}(s^2) = 2(n-1)\sigma^4/n^2 \to 0$ as $n \to \infty$, hence s^2 is consistent for σ^2.

Example 3: For the Bernoulli case, we know that $E(\bar{X}) = \theta$ and $\text{var}(\bar{X}) = \theta(1-\theta)/n \to 0$ as $n \to \infty$, hence $\bar{X}$ is consistent for θ.

Warning: This is only a sufficient condition for consistency. Failing to satisfy this condition does not necessarily imply that the estimator is inconsistent.

(iv) Sufficiency

$\bar{X}$ is sufficient for μ, if $\bar{X}$ contains all the information in the sample pertaining to μ. In other words, $f(X_1, \ldots, X_n/\bar{X})$ is independent of μ. To prove this fact one uses the factorization theorem due to Fisher and Neyman. In this context, $\bar{X}$ is sufficient for μ, if and only if one can factorize the joint p.d.f.

$$f(X_1, \ldots, X_n; \mu) = h(\bar{X}; \mu) \cdot g(X_1, \ldots, X_n)$$

where h and g are any two functions with the latter being only a function of the X's and independent of μ in form and in the domain of the X's.

Example 1: For the Normal case, it is clear from equation (2.1) that by subtracting and adding $\bar{X}$ in the summation we can write after some algebra

$$f(X_1, .., X_n; \mu, \sigma^2) = (1/2\pi\sigma^2)^{n/2} e^{-\{(1/2\sigma^2)\sum_{i=1}^{n}(X_i - \bar{X})^2\}} e^{-\{(n/2\sigma^2)(\bar{X}-\mu)^2\}}$$

Hence, $h(\bar{X}; \mu) = e^{-(n/2\sigma^2)(\bar{X}-\mu^2)}$ and $g(X_1, \ldots, X_n)$ is the remainder term which is independent of μ in form. Also $-\infty < X_i < \infty$ and hence independent of μ in the domain. Therefore, $\bar{X}$ is sufficient for μ.

Example 2: For the Bernoulli case,

$$f(X_1, \ldots, X_n; \theta) = \theta^{n\bar{X}}(1-\theta)^{n(1-\bar{X})} \quad X_i = 0, 1 \quad \text{for } i = 1, \ldots, n.$$

Therefore, $h(\bar{X}, \theta) = \theta^{n\bar{X}}(1-\theta)^{n(1-\bar{X})}$ and $g(X_1, \ldots, X_n) = 1$ which is independent of θ in form and domain. Hence, $\bar{X}$ is sufficient for θ.

Under certain regularity conditions on the distributions we are sampling from, one can show that the MVU of any parameter θ is an unbiased function of a sufficient statistic for θ.[3] Advantages of the maximum likelihood estimators is that (i) they are sufficient estimators when they exist. (ii) They are asymptotically efficient. (iii) If the distribution of the MLE satisfies certain regularity conditions, then making the MLE unbiased results in a unique MVU estimator. A prime example of this is s^2 which was shown to be an unbiased estimator of σ^2 for a random sample drawn from the Normal distribution. It can be shown that s^2 is sufficient for σ^2 and that $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$. Hence, s^2 is an unbiased sufficient statistic for σ^2 and therefore it is MVU for σ^2, even though it does not attain the Cramér-Rao lower bound. (iv) Maximum likelihood estimates are invariant with respect to continuous transformations. To explain the last property, consider the estimator of e^{μ}. Given $\widehat{\mu}_{MLE} = \bar{X}$, an obvious estimator is $e^{\widehat{\mu}_{MLE}} = e^{\bar{X}}$. This is in fact the MLE of e^{μ}. In general, if $g(\mu)$ is a continuous function of μ, then $g(\widehat{\mu}_{MLE})$ is the MLE of $g(\mu)$. Note that $E(e^{\widehat{\mu}_{MLE}}) \neq e^{E(\widehat{\mu}_{MLE})} = e^{\mu}$, in other words, expectations are not invariant to all continuous transformations, especially nonlinear ones and hence the resulting MLE estimator may not be unbiased. $e^{\bar{X}}$ is not unbiased for e^{μ} even though $\bar{X}$ is unbiased for μ.

In summary, there are two routes for finding the MVU estimator. One is systematically following the derivation of a sufficient statistic, proving that its distribution satisfies certain regularity conditions, and then making it unbiased for the parameter in question. Of course, MLE provides us with sufficient statistics, for example,

$$X_1, \ldots, X_n \sim \text{IIN}(\mu, \sigma^2) \Rightarrow \widehat{\mu}_{MLE} = \bar{X} \quad \text{and} \quad \widehat{\sigma}^2_{MLE} = \textstyle\sum_{i=1}^{n}(X_i - \bar{X})^2/n$$

are both sufficient for μ and σ^2, respectively. $\bar{X}$ is unbiased for μ and $\bar{X} \sim N(\mu, \sigma^2/n)$. The Normal distribution satisfies the regularity conditions needed for $\bar{X}$ to be MVU for μ. $\widehat{\sigma}^2_{MLE}$ is biased for σ^2, but $s^2 = n\widehat{\sigma}^2_{MLE}/(n-1)$ is unbiased for σ^2 and $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ which also satisfies the regularity conditions for s^2 to be a MVU estimator for σ^2.

Alternatively, one finds the Cramér-Rao lower bound and checks whether the usual estimator (obtained from say the method of moments or the maximum likelihood method) achieves this lower bound. If it does, this estimator is efficient, and there is no need to search further. If it does not, the former strategy leads us to the MVU estimator. In fact, in the previous example $\bar{X}$ attains the Cramér-Rao lower bound, whereas s^2 does not. However, both are MVU for μ and σ^2 respectively.

(v) Comparing Biased and Unbiased Estimators

Suppose we are given two estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ of θ where the first is unbiased and has a large variance and the second is biased but with a small variance. The question is which one of these two estimators is preferable? $\widehat{\theta}_1$ is unbiased whereas $\widehat{\theta}_2$ is biased. This means that if we repeat the sampling procedure many times then we expect $\widehat{\theta}_1$ to be on the average correct, whereas $\widehat{\theta}_2$ would be on the average different from θ. However, in real life, we observe only one sample. With a large variance for $\widehat{\theta}_1$, there is a great likelihood that the sample drawn could result in a $\widehat{\theta}_1$ far away from θ. However, with a small variance for $\widehat{\theta}_2$, there is a better chance of getting a $\widehat{\theta}_2$ close to θ. If our loss function is $L(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$ then our risk is

$$\begin{aligned} R(\widehat{\theta}, \theta) &= E[L(\widehat{\theta}, \theta)] = E(\widehat{\theta} - \theta)^2 = MSE(\widehat{\theta}) \\ &= E[\widehat{\theta} - E(\widehat{\theta}) + E(\widehat{\theta}) - \theta]^2 = \text{var}(\widehat{\theta}) + (\text{Bias}(\widehat{\theta}))^2. \end{aligned}$$

Minimizing the risk when the loss function is quadratic is equivalent to minimizing the Mean Square Error (MSE). From its definition the MSE shows the trade-off between bias and variance. MVU theory, sets the bias equal to zero and minimizes var$(\widehat{\theta})$. In other words, it minimizes the above risk function but only over $\widehat{\theta}$'s that are unbiased. If we do not restrict ourselves to unbiased estimators of θ, minimizing MSE may result in a biased estimator such as $\widehat{\theta}_2$ which beats $\widehat{\theta}_1$ because the gain from its smaller variance outweighs the loss from its small bias, see Figure 2.2.

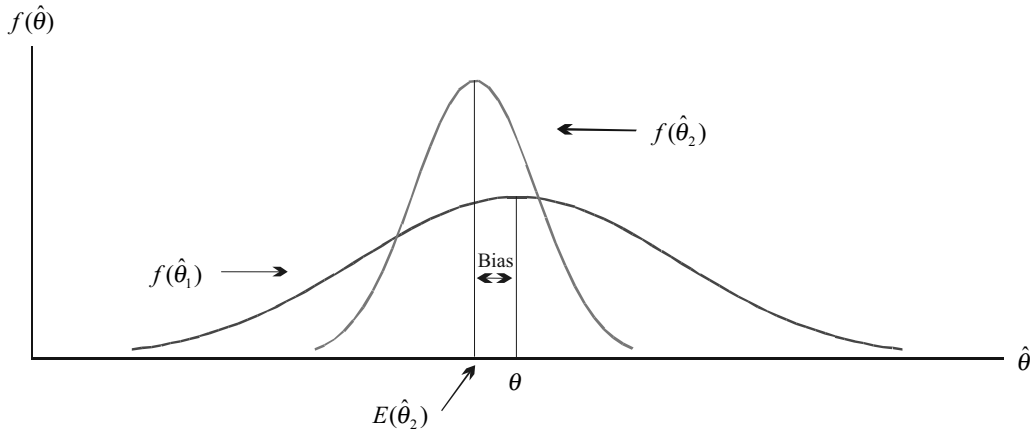


Figure 2.2 Bias versus Variance

2.4 Hypothesis Testing

The best way to proceed is with an example.

Example 1: The Economics Departments instituted a new program to teach micro-principles. We would like to test the null hypothesis that 80% of economics undergraduate students will pass the micro-principles course versus the alternative hypothesis that only 50% will pass. We draw a random sample of size 20 from the large undergraduate micro-principles class and as a simple rule we accept the null if x, the number of passing students is larger or equal to 13, otherwise the alternative hypothesis will be accepted. Note that the distribution we are drawing from is Bernoulli with the probability of success θ, and we have chosen only two states of the world H_0; $\theta_0 = 0.80$ and $H_1; \theta_1 = 0.5$. This situation is known as testing a *simple* hypothesis versus another *simple* hypothesis because the distribution is completely specified under the null or alternative hypothesis. One would expect $(E(x) = n\theta_0)$ 16 students under H_0 and $(n\theta_1)$ 10 students under H_1 to pass the micro-principles exams. It seems then logical to take $x \geq 13$ as the cut-off point distinguishing H_0 form H_1. No theoretical justification is given at this stage to this arbitrary choice except to say that it is the mid-point of $[10, 16]$. Figure 2.3 shows that one can make two types of errors. The first is rejecting H_0 when in fact it is true, this is known as *type I error* and the probability of committing this error is denoted by α. The second is accepting H_1 when it is false. This is known as *type II error* and the corresponding probability is denoted by β. For this example

$$\begin{aligned}\alpha &= \Pr[\text{rejecting } H_0/H_0 \text{ is true}] = \Pr[x < 13/\theta = 0.8] \\ &= b(n=20; x=0; \theta=0.8) + .. + b(n=20; x=12; \theta=0.8) \\ &= b(n=20; x=20; \theta=0.2) + .. + b(n=20; x=8; \theta=0.2) \\ &= 0 + .. + 0 + 0.0001 + 0.0005 + 0.0020 + 0.0074 + 0.0222 = 0.0322\end{aligned}$$

where we have used the fact that $b(n; x; \theta) = b(n; n-x; 1-\theta)$ and $b(n; x; \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ denotes the binomial distribution for $x = 0, 1, \ldots, n$, see problem 4.

		True World	
		$\theta_0 = 0.80$	$\theta_1 = 0.50$
Decision	θ_0	No error	Type II error
	θ_1	Type I error	No Error

Figure 2.3 Type I and II Error

$$\begin{aligned}\beta &= \Pr[\text{accepting } H_0/H_0 \text{ is false}] = \Pr[x \geq 13/\theta = 0.5] \\ &= b(n=20; x=13; \theta=0.5) + .. + b(n=20; x=20; \theta=0.5) \\ &= 0.0739 + 0.0370 + 0.0148 + 0.0046 + 0.0011 + 0.0002 + 0 + 0 = 0.1316\end{aligned}$$

The rejection region for $H_0, x < 13$, is known as the *critical region* of the test and $\alpha = \Pr[\text{Falling in the critical region}/H_0 \text{ is true}]$ is also known as the *size* of the critical region. A good test is one which minimizes both types of errors α and β. For the above example, α is low but β is high with more than a 13% chance of happening. This β can be reduced by changing the critical region from $x < 13$ to $x < 14$, so that H_0 is accepted only if $x \geq 14$. In this case, one can easily verify that

$$\begin{aligned}\alpha &= \Pr[x < 14/\theta = 0.8] = b(n=20;\ x=0; \theta=0.8) + .. + b(n=20, x=13, \theta=0.8) \\ &= 0.0322 + b(n=20; x=13;\ \theta=0.8) = 0.0322 + 0.0545 = 0.0867\end{aligned}$$

and

$$\begin{aligned}\beta &= \Pr[x \geq 14/\theta = 0.5] = b(n=20;\ x=14; \theta=0.5) + .. + b(n=20; x=20; \theta=0.5) \\ &= 0.1316 - b(n=20; x=13;\ \theta=0.5) = 0.0577\end{aligned}$$

By becoming more conservative on accepting H_0 and more liberal on accepting H_1, one reduces β from 0.1316 to 0.0577 but the price paid is the increase in α from 0.0322 to 0.0867. The only way to reduce both α and β is by increasing n. For a fixed n, there is a tradeoff between α and β as we change the critical region. To understand this clearly, consider the real life situation of trial by jury for which the defendant can be innocent or guilty. The decision of incarceration or release implies two types of errors. One can make $\alpha = \Pr[\text{incarcerating/innocence}] = 0$ and $\beta =$ its maximum, by releasing *every* defendant. Or one can make $\beta = \Pr[\text{release/guilty}] = 0$ and $\alpha =$ its maximum, by incarcerating *every* defendant. These are extreme cases but hopefully they demonstrate the trade-off between α and β.

The Neyman-Pearson Theory

The classical theory of hypothesis testing, known as the Neyman-Pearson theory, fixes $\alpha = \Pr(\text{type I error}) \leq$ a constant and minimizes β or maximizes $(1-\beta)$. The latter is known as the *Power* of the test under the alternative.

The Neyman-Pearson Lemma: If C is a critical region of size α and k is a constant such that

$$(L_0/L_1) \leq k \text{ inside } C$$

and

$$(L_0/L_1) \geq k \text{ outside } C$$

then C is a most powerful critical region of size α for testing $H_0; \theta = \theta_0$, against $H_1; \theta = \theta_1$.

Note that the likelihood has to be completely specified under the null and alternative. Hence, this lemma applies only to testing a simple versus another simple hypothesis. The proof of this lemma is given in Freund (1992). Intuitively, L_0 is the likelihood function under the null H_0 and L_1 is the corresponding likelihood function under H_1. Therefore, (L_0/L_1) should be small for points inside the critical region C and large for points outside the critical region C. The proof of the theorem shows that any other critical region, say D, of size α cannot have a smaller probability of type II error than C. Therefore, C is the best or most powerful critical region of size α. Its power $(1-\beta)$ is maximum at H_1. Let us demonstrate this lemma with an example.

Example 2: Given a random sample of size n from $N(\mu, \sigma^2 = 4)$, use the Neyman-Pearson lemma to find the most powerful critical region of size $\alpha = 0.05$ for testing $H_0; \mu_0 = 2$ against the alternative $H_1; \mu_1 = 4$.

Note that this is a simple versus simple hypothesis as required by the lemma, since $\sigma^2 = 4$ is known and μ is specified by H_0 and H_1. The likelihood function for the $N(\mu, 4)$ density is given by

$$L(\mu) = f(x_1, \ldots, x_n; \mu, 4) = (1/2\sqrt{2\pi})^n \exp\left\{-\textstyle\sum_{i=1}^n (x_i - \mu)^2/8\right\}$$

so that

$$L_0 = L(\mu_0) = (1/2\sqrt{2\pi})^n \exp\left\{-\textstyle\sum_{i=1}^n (x_i - 2)^2/8\right\}$$

and

$$L_1 = L(\mu_1) = (1/2\sqrt{2\pi})^n \exp\left\{-\textstyle\sum_{i=1}^n (x_i - 4)^2/8\right\}$$

Therefore

$$L_0/L_1 = \exp\left\{-\left[\textstyle\sum_{i=1}^n (x_i - 2)^2 - \sum_{i=1}^n (x_i - 4)^2\right]/8\right\} = \exp\left\{-\textstyle\sum_{i=1}^n x_i/2 + 3n/2\right\}$$

and the critical region is defined by

$$\exp\left\{-\textstyle\sum_{i=1}^n x_i/2 + 3n/2\right\} \leq k \quad \text{inside } C$$

Taking logarithms of both sides, subtracting $(3/2)n$ and dividing by $(-1/2)n$ one gets

$$\bar{x} \geq K \quad \text{inside } C$$

In practice, one need not keep track of K as long as one keeps track of the direction of the inequality. K can be determined by making the size of $C = \alpha = 0.05$. In this case

$$\alpha = \Pr[\bar{x} \geq K/\mu = 2] = \Pr[z \geq (K-2)/(2/\sqrt{n})]$$

where $z = (\bar{x} - 2)/(2/\sqrt{n})$ is distributed $N(0,1)$ under H_0. From the $N(0,1)$ tables, we have

$$\frac{K-2}{(2/\sqrt{n})} = 1.645$$

Hence,

$$K = 2 + 1.645(2/\sqrt{n})$$

and $\bar{x} \geq 2 + 1.645(2/\sqrt{n})$ defines the most powerful critical region of size $\alpha = 0.05$ for testing $H_0; \mu_0 = 2$ versus $H_1; \mu_1 = 4$. Note that, in this case

$$\begin{aligned} \beta &= \Pr[\bar{x} < 2 + 1.645(2/\sqrt{n})/\mu = 4] \\ &= \Pr[z < [-2 + 1.645(2/\sqrt{n})]/(2/\sqrt{n})] = \Pr[z < 1.645 - \sqrt{n}] \end{aligned}$$

For $n = 4$; $\beta = \Pr[z < -0.355] = 0.3613$ shown by the shaded region in Figure 2.4. For $n = 9$; $\beta = \Pr[z < -1.355] = 0.0877$, and for $n = 16$; $\beta = \Pr[z < -2.355] = 0.00925$.

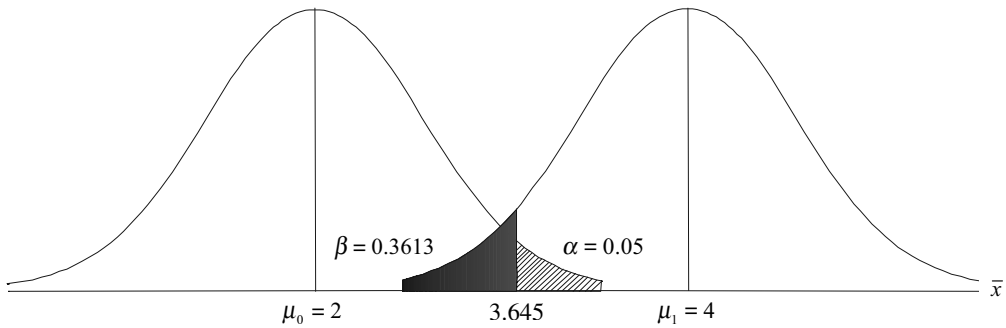


Figure 2.4 Critical Region for Testing $\mu_0 = 2$ against $\mu_1 = 4$ for $n = 4$

This gives us an idea of how, for a fixed $\alpha = 0.05$, the minimum β decreases with larger sample size n. As n increases from 4 to 9 to 16, the $\text{var}(\bar{x}) = \sigma^2/n$ decreases and the two distributions shown in Figure 2.4 shrink in dispersion still centered around $\mu_0 = 2$ and $\mu_1 = 4$, respectively. This allows better decision making (based on larger sample size) as reflected by the critical region shrinking from $\overline{x} \geq 3.65$ for $n = 4$ to $\overline{x} \geq 2.8225$ for $n = 16$, and the power $(1 - \beta)$ rising from 0.6387 to 0.9908, respectively, for a fixed $\alpha \leq 0.05$. The power function is the probability of rejecting H_0. It is equal to α under H_0 and $1 - \beta$ under H_1. The ideal power function is zero at H_0 and one at H_1. The Neyman-Pearson lemma allows us to fix α, say at 0.05, and find the test with the best power at H_1.

In example 2, both the null and alternative hypotheses are simple. In real life, one is more likely to be faced with testing $H_0; \mu = 2$ versus $H_1; \mu \neq 2$. Under the alternative hypothesis, the distribution is not completely specified, since the mean μ is not known, and this is referred to as a *composite* hypothesis. In this case, one cannot compute the probability of type II error

since the distribution is not known under the alternative. Also, the Neyman-Pearson lemma cannot be applied. However, a simple generalization allows us to compute a Likelihood Ratio test which has satisfactory properties but is no longer uniformly most powerful of size α. In this case, one replaces L_1, which is not known since H_1 is a composite hypothesis, by the maximum value of the likelihood, i.e.,

$$\lambda = \frac{\max L_0}{\max L}$$

Since max L_0 is the maximum value of the likelihood under the null while max$L$ is the maximum value of the likelihood over the whole parameter space, it follows that max$L_0 \leq$ max$L$ and $\lambda \leq 1$. Hence, if H_0 is true, λ is close to 1, otherwise it is smaller than 1. Therefore, $\lambda \leq k$ defines the critical region for the Likelihood Ratio test, and k is determined such that the size of this test is α.

Example 3: For a random sample $x_1, \ldots, x_n$ drawn from a Normal distribution with mean μ and variance $\sigma^2 = 4$, derive the Likelihood Ratio test for $H_0; \mu = 2$ versus $H_1; \mu \neq 2$. In this case

$$\max L_0 = (1/2\sqrt{2\pi})^n \exp\left\{-\textstyle\sum_{i=1}^n (x_i - 2)^2/8\right\} = L_0$$

and

$$\max L = (1/2\sqrt{2\pi})^n \exp\left\{-\textstyle\sum_{i=1}^n (x_i - \bar{x})^2/8\right\} = L(\widehat{\mu}_{MLE})$$

where use is made of the fact that $\widehat{\mu}_{MLE} = \bar{x}$. Therefore,

$$\lambda = \exp\left\{\left[-\textstyle\sum_{i=1}^n (x_i - 2)^2 + \sum_{i=1}^n (x_i - \bar{x})^2\right]/8\right\} = \exp\left\{-n(\bar{x} - 2)^2/8\right\}$$

Hence, the region for which $\lambda \leq k$, is equivalent after some simple algebra to the following region

$$(\bar{x} - 2)^2 \geq K \quad \text{or} \quad |\bar{x} - 2| \geq K^{1/2}$$

where K is determined such that

$$\Pr[|\bar{x} - 2| \geq K^{1/2}/\mu = 2] = \alpha$$

We know that $\bar{x} \sim N(2, 4/n)$ under H_0. Hence, $z = (\bar{x} - 2)/(2/\sqrt{n})$ is $N(0,1)$ under H_0, and the critical region of size α will be based upon $|z| \geq z_{\alpha/2}$ where $z_{\alpha/2}$ is given in Figure 2.5 and is the value of a $N(0,1)$ random variable such that the probability of exceeding it is $\alpha/2$. For $\alpha = 0.05$, $z_{\alpha/2} = 1.96$, and for $\alpha = 0.10$, $z_{\alpha/2} = 1.645$. This is a two-tailed test with rejection of H_0 obtained in case $z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$.

Note that in this case

$$LR \equiv -2\log\lambda = (\bar{x} - 2)^2/(4/n) = z^2$$

which is distributed as χ^2_1 under H_0. This is because it is the square of a $N(0,1)$ random variable under H_0. This is a finite sample result holding for any n. In general, other examples may lead to more complicated λ statistics for which it is difficult to find the corresponding distributions and hence the corresponding critical values. For these cases, we have an asymptotic result

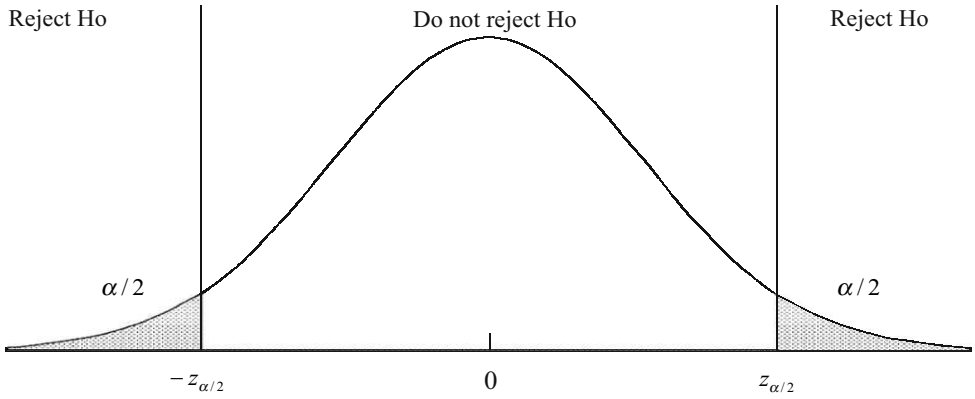


Figure 2.5 Critical Values

which states that, for large n, $LR = -2\log\lambda$ will be asymptotically distributed as χ^2_ν where ν denotes the number of restrictions that are tested by H_0. For example 2, $\nu = 1$ and hence, LR is asymptotically distributed as χ^2_1. Note that we did not need this result as we found LR is exactly distributed as χ^2_1 for any n. If one is testing H_0; $\mu = 2$ and $\sigma^2 = 4$ against the alternative that H_1; $\mu \neq 2$ or $\sigma^2 \neq 4$, then the corresponding LR will be asymptotically distributed as χ^2_2, see problem 5, part (f).

Likelihood Ratio, Wald and Lagrange Multiplier Tests

Before we go into the derivations of these three tests we start by giving an intuitive graphical explanation that will hopefully emphasize the differences among these tests. This intuitive explanation is based on the article by Buse (1982).

Consider a quadratic log-likelihood function in a parameter of interest, say μ. Figure 2.6 shows this log-likelihood $\log L(\mu)$, with a maximum at $\widehat{\mu}$. The Likelihood Ratio test, tests the null hypothesis $H_0; \mu = \mu_0$ by looking at the ratio of the likelihoods $\lambda = L(\mu_0)/L(\widehat{\mu})$ where

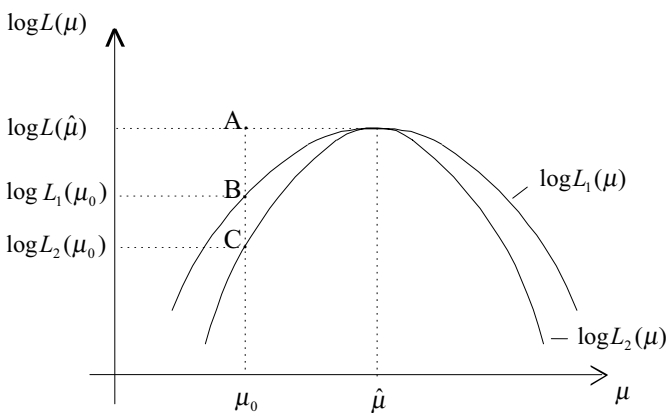


Figure 2.6 Wald Test

$-2\log\lambda$, twice the difference in log-likelihood, is distributed asymptotically as χ^2_1 under H_0. This test differentiates between the top of the hill and a preassigned point on the hill by evaluating the height at both points. Therefore, it needs both the restricted and unrestricted maximum of the likelihood. This ratio is dependent on the distance of μ_0 from $\widehat{\mu}$ and the curvature of the log-likelihood, $C(\mu) = |\partial^2 \log L(\mu)/\partial\mu^2|$, at $\widehat{\mu}$. In fact, for a fixed $(\widehat{\mu} - \mu_0)$, the larger $C(\widehat{\mu})$, the larger is the difference between the two heights. Also, for a given curvature at $\widehat{\mu}$, the larger $(\widehat{\mu} - \mu_0)$ the larger is the difference between the heights. The Wald test works from the top of the hill, i.e., it needs only the unrestricted maximum likelihood. It tries to establish the distance to μ_0, by looking at the horizontal distance $(\widehat{\mu} - \mu_0)$, and the curvature at $\widehat{\mu}$. In fact the Wald statistic is $W = (\widehat{\mu} - \mu_0)^2 C(\widehat{\mu})$ and this is asymptotically distributed as χ^2_1 under H_0. The usual form of W has $I(\mu) = -E[\partial^2 \log L(\mu)/\partial\mu^2]$ the information matrix evaluated at $\widehat{\mu}$, rather than $C(\widehat{\mu})$, but the latter is a consistent estimator of $I(\mu)$. The information matrix will be studied in details in Chapter 7. It will be shown, under fairly general conditions, that $\widehat{\mu}$ the MLE of μ, has $\text{var}(\widehat{\mu}) = I^{-1}(\mu)$. Hence $W = (\widehat{\mu} - \mu_0)^2/\text{var}(\widehat{\mu})$ all evaluated at the unrestricted MLE. The Lagrange-Multiplier test (LM), on the other hand, goes to the preassigned point μ_0, i.e., it only needs the restricted maximum likelihood, and tries to determine how far it is from the top of the hill by considering the slope of the tangent to the likelihood $S(\mu) = \partial \log L(\mu)/\partial\mu$ at μ_0, and the rate at which this slope is changing, i.e., the curvature at μ_0. As Figure 2.7 shows, for two log-likelihoods with the same $S(\mu_0)$, the one that is closer to the top of the hill is the one with the larger curvature at μ_0.

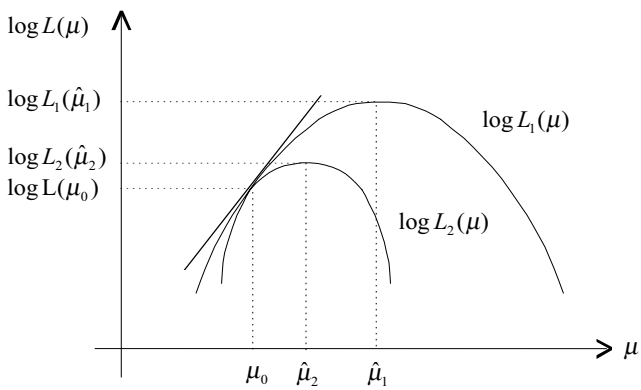


Figure 2.7 LM Test

This suggests the following statistic: $LM = S^2(\mu_0)\{C(\mu_0)\}^{-1}$ where the curvature appears in inverse form. In the Appendix to this chapter, we show that the $E[S(\mu)] = 0$ and $\text{var}[S(\mu)] = I(\mu)$. Hence $LM = S^2(\mu_0)I^{-1}(\mu_0) = S^2(\mu_0)/\text{var}[S(\mu_0)]$ all evaluated at the restricted MLE. Another interpretation of the LM test is that it is a measure of failure of the restricted estimator, in this case μ_0, to satisfy the first-order conditions of maximization of the unrestricted likelihood. We know that $S(\widehat{\mu}) = 0$. The question is: to what extent does $S(\mu_0)$ differ from zero? $S(\mu)$ is known in the statistics literature as the *score*, and the LM test is also referred to as the score test. For a more formal treatment of these tests, let us reconsider example 3 of a random sample $x_1, \ldots, x_n$ from a $N(\mu, 4)$ where we are interested in testing $H_0; \mu_0 = 2$ versus $H_1; \mu \neq 2$. The likelihood function $L(\mu)$ as well as $LR = -2\log\lambda = n(\bar{x} - 2)^2/4$ were given in example 3. In

fact, the score function is given by

$$S(\mu) = \frac{\partial \log L(\mu)}{\partial \mu} = \frac{\sum_{i=1}^{n}(x_i - \mu)}{4} = \frac{n(\bar{x} - \mu)}{4}$$

and under H_0

$$\begin{aligned} S(\mu_0) &= S(2) = \frac{n(\bar{x} - 2)}{4} \\ C(\mu) &= |\frac{\partial^2 \log L(\mu)}{\partial \mu^2}| = |-\frac{n}{4}| = \frac{n}{4} \end{aligned}$$

and $I(\mu) = -E\left[\frac{\partial^2 \log L(\mu)}{\partial \mu^2}\right] = \frac{n}{4} = C(\mu)$.

The Wald statistic is based on

$$W = (\widehat{\mu}_{MLE} - 2)^2 I(\widehat{\mu}_{MLE}) = (\bar{x} - 2)^2 \cdot \left(\frac{n}{4}\right)$$

The LM statistic is based on

$$LM = S^2(\mu_0) I^{-1}(\mu_0) = \frac{n^2(\bar{x} - 2)^2}{16} \cdot \frac{4}{n} = \frac{n(\bar{x} - 2)^2}{4}$$

Therefore, $W = LM = LR$ for this example with known variance $\sigma^2 = 4$. These tests are all based upon the $|\bar{x} - 2| \geq k$ critical region, where k is determined such that the size of the test is α. In general, these test statistics are not always equal, as is shown in the next example.

Example 4: For a random sample $x_1, \ldots, x_n$ drawn from a $N(\mu, \sigma^2)$ with *unknown* σ^2, test the hypothesis $H_0; \mu = 2$ versus $H_1; \mu \neq 2$. Problem 5, part (c), asks the reader to verify that

$$LR = n \log\left[\frac{\sum_{i=1}^{n}(x_i - 2)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \text{ whereas } W = \frac{n^2(\bar{x} - 2)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ and } LM = \frac{n^2(\bar{x} - 2)^2}{\sum_{i=1}^{n}(x_i - 2)^2}.$$

One can easily show that $LM/n = (W/n)/[1+(W/n)]$ and $LR/n = \log[1+(W/n)]$. Let $y = W/n$, then using the inequality $y \geq \log(1 + y) \geq y/(1 + y)$, one can conclude that $W \geq LR \geq LM$. This inequality was derived by Berndt and Savin (1977), and will be considered again when we study test of hypotheses in the general linear model. Note, however that all three test statistics are based upon $|\bar{x} - 2| \geq k$ and for finite n, the same exact critical value could be obtained from the Normally distributed $\bar{x}$. This section introduced the W, LR and LM test statistics, all of which have the same asymptotic distribution. In addition, we showed that using the normal distribution, when σ^2 is known, $W = LR = LM$ for testing $H_0; \mu = 2$ versus $H_1; \mu \neq 2$. However, when σ^2 is unknown, we showed that $W \geq LR \geq LM$ for the same hypothesis.

Example 5: For a random sample $x_1, \ldots, x_n$ drawn from a Bernoulli distribution with parameter θ, test the hypothesis $H_0; \theta = \theta_0$ versus $H_1; \theta \neq \theta_0$, where θ_0 is a known positive fraction. This example is based on Engle (1984). Problem 4, part (i), asks the reader to derive LR, W and LM for H_0; $\theta = 0.2$ versus H_1; $\theta \neq 0.2$. The likelihood $L(\theta)$ and the Score $S(\theta)$ were derived in section 2.2. One can easily verify that

$$C(\theta) = |\frac{\partial^2 \log L(\theta)}{\partial \theta^2}| = \frac{\sum_{i=1}^{n} x_i}{\theta^2} + \frac{n - \sum_{i=1}^{n} x_i}{(1 - \theta)^2}$$

and

$$I(\theta) = -E\left[\frac{\partial^2 \log L(\theta)}{\partial \theta^2}\right] = \frac{n}{\theta(1-\theta)}$$

The Wald statistic is based on

$$W = (\widehat{\theta}_{MLE} - \theta_0)^2 I(\widehat{\theta}_{MLE}) = (\bar{x} - \theta_0)^2 \cdot \frac{n}{\bar{x}(1-\bar{x})} = \frac{(\bar{x} - \theta_0)^2}{\bar{x}(1-\bar{x})/n}$$

using the fact that $\widehat{\theta}_{MLE} = \bar{x}$. The LM statistic is based on

$$LM = S^2(\theta_0) I^{-1}(\theta_0) = \frac{(\bar{x} - \theta_0)^2}{[\theta_0(1-\theta_0)/n]^2} \cdot \frac{\theta_0(1-\theta_0)}{n} = \frac{(\bar{x} - \theta_0)^2}{\theta_0(1-\theta_0)/n}$$

Note that the numerator of the W and LM are the same. It is the denominator which is the $\text{var}(\bar{x}) = \theta(1-\theta)/n$ that is different. For Wald, this $\text{var}(\bar{x})$ is evaluated at $\widehat{\theta}_{MLE}$, whereas for LM, this is evaluated at θ_0.

The LR statistic is based on

$$\log L(\widehat{\theta}_{MLE}) = \textstyle\sum_{i=1}^n x_i \log \bar{x} + (n - \sum_{i=1}^n x_i) \log(1 - \bar{x})$$

and

$$\log L(\theta_0) = \textstyle\sum_{i=1}^n x_i \log \theta_0 + (n - \sum_{i=1}^n x_i) \log(1 - \theta_0)$$

so that

$$\begin{aligned} LR &= -2\log L(\theta_0) + 2\log L(\widehat{\theta}_{MLE}) = -2[\textstyle\sum_{i=1}^n x_i (\log \theta_0 - \log \bar{x}) \\ &\quad + (n - \sum_{i=1}^n x_i)(\log(1-\theta_0) - \log(1-\bar{x}))] \end{aligned}$$

For this example, LR looks different from W and LM. However, a second-order Taylor-Series expansion of LR around $\theta_0 = \bar{x}$ yields the same statistic. Also, for $n \to \infty$, plim $\bar{x} = \theta$ and if H_0 is true, then all three statistics are asymptotically equivalent. Note also that all three test statistics are based upon $|\bar{x} - \theta_0| \geq k$ and for finite n, the same exact critical value could be obtained from the binomial distribution. See problem 19 for more examples of the conflict in test of hypotheses using the W, LR and LM test statistics.

Bera and Permaratne (2001, p. 58) tell the following amusing story that can bring home the interrelationship among the three tests: "Once around 1946 Ronald Fisher invited Jerzy Neyman, Abraham Wald, and C.R. Rao to his lodge for afternoon tea. During their conversation, Fisher mentioned the problem of deciding whether his dog, who had been going to an "obedience school" for some time, was disciplined enough. Neyman quickly came up with an idea: leave the dog free for some time and then put him on his leash. If there is not much difference in his behavior, the dog can be thought of as having completed the course successfully. Wald, who lost his family in the concentration camps, was adverse to any restrictions and simply suggested leaving the dog free and seeing whether it behaved properly. Rao, who had observed the nuisances of stray dogs in Calcutta streets did not like the idea of letting the dog roam freely and suggested keeping the dog on a leash at all times and observing how hard it pulls on the leash. If it pulled too much, it needed more training. That night when Rao was back in his Cambridge dormitory after tending Fisher's mice at the genetics laboratory, he suddenly realized the connection of Neyman and Wald's recommendations to the Neyman-Pearson LR and Wald tests. He got an idea and the rest is history."

2.5 Confidence Intervals

Estimation methods considered in section 2.2 give us a point estimate of a parameter, say μ, and that is the best bet, given the data and the estimation method, of what μ might be. But it is always good policy to give the client an interval, rather than a point estimate, where with some degree of confidence, usually 95% confidence, we expect μ to lie. We have seen in Figure 2.5 that for a $N(0,1)$ random variable z, we have

$$\Pr[-z_{\alpha/2} \leq z \leq z_{\alpha/2}] = 1 - \alpha$$

and for $\alpha = 5\%$, this probability is 0.95, giving the required 95% confidence. In fact, $z_{\alpha/2} = 1.96$ and

$$\Pr[-1.96 \leq z \leq 1.96] = 0.95$$

This says that if we draw 100 random numbers from a $N(0,1)$ density, (using a normal random number generator) we expect 95 out of these 100 numbers to lie in the $[-1.96, 1.96]$ interval. Now, let us get back to the problem of estimating μ from a random sample $x_1, \ldots, x_n$ drawn from a $N(\mu, \sigma^2)$ distribution. We found out that $\widehat{\mu}_{MLE} = \bar{x}$ and $\bar{x} \sim N(\mu, \sigma^2/n)$. Hence, $z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$ is $N(0,1)$. The point estimate for μ is $\bar{x}$ observed from the sample, and the 95% confidence interval for μ is obtained by replacing z by its value in the above probability statement:

$$\Pr[-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}] = 1 - \alpha$$

Assuming σ is known for the moment, one can rewrite this probability statement after some simple algebraic manipulations as

$$\Pr[\bar{x} - z_{\alpha/2}(\sigma/\sqrt{n}) \leq \mu \leq \bar{x} + z_{\alpha/2}(\sigma/\sqrt{n})] = 1 - \alpha$$

Note that this probability statement has random variables on both ends and the probability that these random variables sandwich the unknown parameter μ is $1 - \alpha$. With the same confidence of drawing 100 random $N(0,1)$ numbers and finding 95 of them falling in the $(-1.96, 1.96)$ range we are confident that if we drew a 100 samples and computed a 100 $\bar{x}$'s, and a 100 intervals $(\bar{x} \pm 1.96\ \sigma/\sqrt{n})$, μ will lie in these intervals in 95 out of 100 times.

If σ is not known, and is replaced by s, then problem 12 shows that this is equivalent to dividing a $N(0,1)$ random variable by an independent χ^2_{n-1} random variable divided by its degrees of freedom, leading to a t-distribution with $(n-1)$ degrees of freedom. Hence, using the t-tables for $(n-1)$ degrees of freedom

$$\Pr[-t_{\alpha/2;n-1} \leq t_{n-1} \leq t_{\alpha/2;n-1}] = 1 - \alpha$$

and replacing t_{n-1} by $(\bar{x} - \mu)/(s/\sqrt{n})$ one gets

$$\Pr[\bar{x} - t_{\alpha/2;n-1}(s/\sqrt{n}) \leq \mu \leq \bar{x} + t_{\alpha/2;n-1}(s/\sqrt{n})] = 1 - \alpha$$

Note that the degrees of freedom $(n-1)$ for the t-distribution come from s and the corresponding critical value $t_{n-1;\alpha/2}$ is therefore sample specific, unlike the corresponding case for the normal density where $z_{\alpha/2}$ does not depend on n. For small n, the $t_{\alpha/2}$ values differ drastically from

Table 2.1 Descriptive Statistics for the Earnings Data

Sample: 1 595								
	LWAGE	WKS	ED	EX	MS	FEM	BLK	UNION
Mean	6.9507	46.4520	12.8450	22.8540	0.8050	0.1126	0.0723	0.3664
Median	6.9847	48.0000	12.0000	21.0000	1.0000	0.0000	0.0000	0.0000
Maximum	8.5370	52.0000	17.0000	51.0000	1.0000	1.0000	1.0000	1.0000
Minimum	5.6768	5.0000	4.0000	7.0000	0.0000	0.0000	0.0000	0.0000
Std. Dev.	0.4384	5.1850	2.7900	10.7900	0.3965	0.3164	0.2592	0.4822
Skewness	–0.1140	–2.7309	–0.2581	0.4208	–1.5400	2.4510	3.3038	0.5546
Kurtosis	3.3937	13.7780	2.7127	2.0086	3.3715	7.0075	11.9150	1.3076
Jarque-Bera	5.13	3619.40	8.65	41.93	238.59	993.90	3052.80	101.51
Probability	0.0769	0.0000	0.0132	0.0000	0.0000	0.0000	0.0000	0.0000
Observations	595	595	595	595	595	595	595	595

$z_{\alpha/2}$, emphasizing the importance of using the t-density in small samples. When n is large the difference between $z_{\alpha/2}$ and $t_{\alpha/2}$ diminishes as the t-density becomes more like a normal density. For $n = 20$, and $\alpha = 0.05, t_{\alpha/2;n-1} = 2.093$ as compared with $z_{\alpha/2} = 1.96$. Therefore,

$$\Pr[-2.093 \leq t_{n-1} \leq 2.093] = 0.95$$

and μ lies in $\bar{x} \pm 2.093(s/\sqrt{n})$ with 95% confidence.

More examples of confidence intervals can be constructed, but the idea should be clear. Note that these confidence intervals are the other side of the coin for tests of hypotheses. For example, in testing $H_0; \mu = 2$ versus $H_1; \mu \neq 2$ for a known σ, we discovered that the Likelihood Ratio test is based on the same probability statement that generated the confidence interval for μ. In classical tests of hypothesis, we choose the level of confidence $\alpha = 5\%$ and compute $z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$. This can be done since σ is known and $\mu = 2$ under the null hypothesis H_0. Next, we do not reject H_0 if z lies in the $(-z_{\alpha/2}, z_{\alpha/2})$ interval and reject H_0 otherwise. For confidence intervals, on the other hand, we do not know μ, and armed with a level of confidence $(1 - \alpha)\%$ we construct the interval that should contain μ with that level of confidence. Having done that, if $\mu = 2$ lies in that 95% confidence interval, then we cannot reject $H_0; \mu = 2$ at the 5% level. Otherwise, we reject H_0. This highlights the fact that any value of μ that lies in this 95% confidence interval (assuming it was our null hypothesis) cannot be rejected at the 5% level by this sample. This is why we do not say "accept H_0", but rather we say "do not reject H_0".

2.6 Descriptive Statistics

In Chapter 4, we will consider the estimation of a simple wage equation based on 595 individuals drawn from the Panel Study of Income Dynamics for 1982. This data is available on the Springer web site as EARN.ASC. Table 2.1 gives the descriptive statistics using EViews for a subset of the variables in this data set.

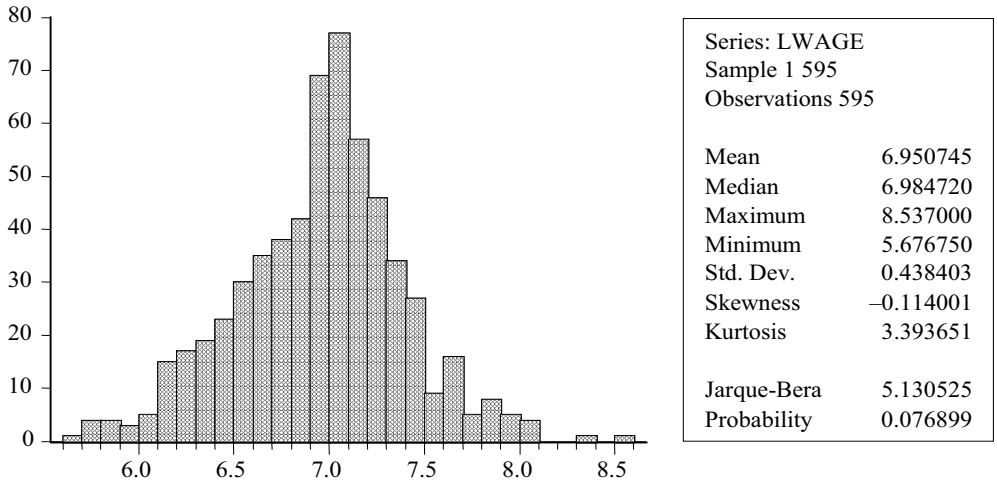


Figure 2.8 Log (Wage) Histogram

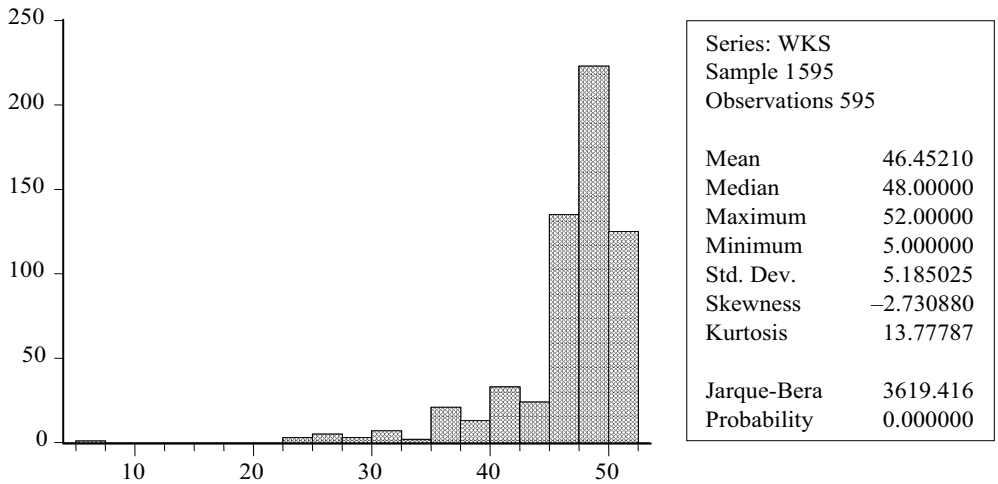


Figure 2.9 Weeks Worked Histogram

The average log wage is \$6.95 for this sample with a minimum of \$5.68 and a maximum of \$8.54. The standard deviation of log wage is 0.44. A plot of the log wage histogram is given in Figure 2.8. Weeks worked vary between 5 and 52 with an average of 46.5 and a standard deviation of 5.2. This variable is highly skewed as evidenced by the histogram in Figure 2.9. Years of education vary between 4 and 17 with an average of 12.8 and a standard deviation of 2.79. There is the usual bunching up at 12 years, which is also the median, as is clear from Figure 2.10.

Experience varies between 7 and 51 with an average of 22.9 and a standard deviation of 10.79. The distribution of this variable is skewed to the left, as shown in Figure 2.11.

Marital status is a qualitative variable indicating whether the individual is married or not. This information is recoded as a numeric $(1, 0)$ variable, one if the individual is married and zero

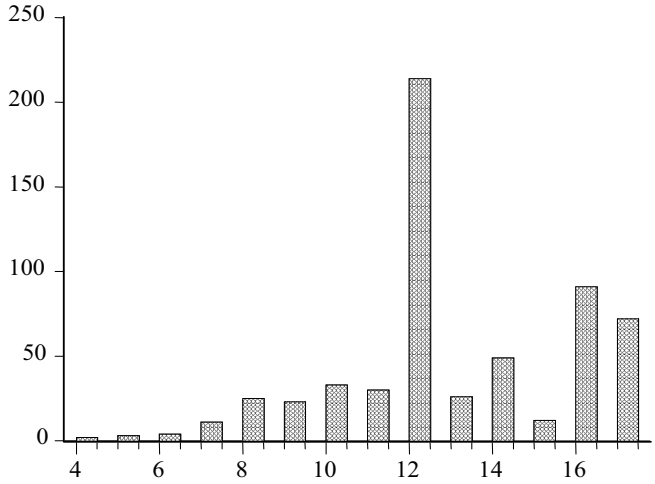


Figure 2.10 Years of Education Histogram

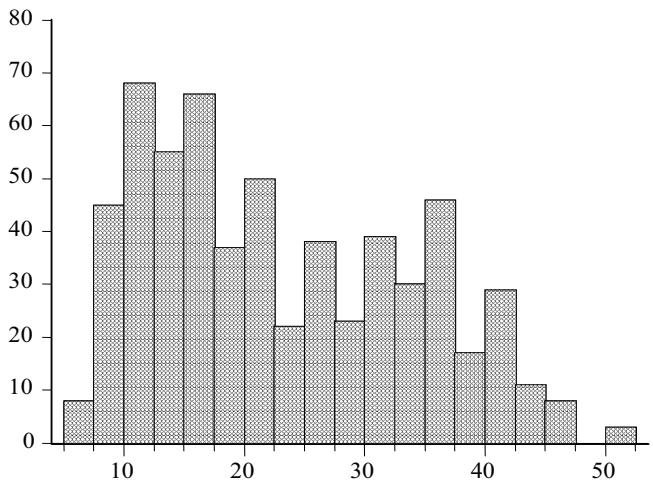


Figure 2.11 Years of Experience Histogram

otherwise. This recoded variable is also known as a dummy variable. It is basically a switch turning on when the individual is married and off when he or she is not. Female is another dummy variable taking the value one when the individual is a female and zero otherwise. Black is a dummy variable taking the value one when the individual is black and zero otherwise. Union is a dummy variable taking the value one if the individual's wage is set by a union contract and zero otherwise. The minimum and maximum values for these dummy variables are obvious. But if they were not zero and one, respectively, you know that something is wrong. The average is a meaningful statistic indicating the percentage of married individuals, females, blacks and union contracted wages in the sample. These are 80.5, 11.3, 7.2 and 30.6%, respectively. We would like to investigate the following claims: (i) women are paid less than men; (ii) blacks are paid less than non-blacks; (iii) married individuals earn more than non-married individuals; and (iv) union contracted wages are higher than non-union wages.

Table 2.2 Test for the Difference in Means

	Average log wage	Difference
Male	$7,004	–0.474
Female	$6,530	(–8.86)
Non-Black	$6,978	–0.377
Black	$6,601	(–5.57)
Not Married	$6,664	0.356
Married	$7,020	(8.28)
Non-Union	$6,945	0.017
Union	$6,962	(0.45)

Table 2.3 Correlation Matrix

	LWAGE	WKS	ED	EX	MS	FEM	BLK	UNION
LWAGE	1.0000	0.0403	0.4566	0.0873	0.3218	–0.3419	–0.2229	0.0183
WKS	0.0403	1.0000	0.0002	–0.1061	0.0782	–0.0875	–0.0594	–0.1721
ED	0.4566	0.0002	1.0000	–0.2219	0.0184	–0.0012	–0.1196	–0.2719
EX	0.0873	–0.1061	–0.2219	1.0000	0.1570	–0.0938	0.0411	0.0689
MS	0.3218	0.0782	0.0184	0.1570	1.0000	–0.7104	–0.2231	0.1189
FEM	–0.3419	–0.0875	–0.0012	–0.0938	–0.7104	1.0000	0.2086	–0.1274
BLK	–0.2229	–0.0594	–0.1196	0.0411	–0.2231	0.2086	1.0000	0.0302
UNION	0.0183	–0.1721	–0.2719	0.0689	0.1189	–0.1274	0.0302	1.0000

A simple first check could be based on computing the average log wage for each of these categories and testing whether the difference in means is significantly different from zero. This can be done using a t-test, see Table 2.2. The average log wage for males and females is given along with their difference and the corresponding t-statistic for the significance of this difference. Other rows of Table 2.2 give similar statistics for other groupings. In Chapter 4, we will show that this t-test can be obtained from a simple regression of log wage on the categorical dummy variable distinguishing the two groups. In this case, the Female dummy variable. From Table 2.2, it is clear that only the difference between union and non-union contracted wages are insignificant.

One can also plot log wage versus experience, see Figure 2.12, log wage versus education, see Figure 2.13, and log wage versus weeks, see Figure 2.14.

The data shows that, in general, log wage increases with education level, weeks worked, but that it exhibits a rising and then a declining pattern with more years of experience. Note that the t-tests based on the difference in log wage across two groupings of individuals, by sex, race or marital status, or the figures plotting log wage versus education, log wage versus weeks worked are based on pairs of variables in each case. A nice summary statistic based also on pairwise comparisons of these variables is the correlation matrix across the data. This is given in Table 2.3.

The signs of this correlation matrix give the direction of linear relationship between the corresponding two variables, while the magnitude gives the strength of this correlation. In Chapter 3, we will see that these simple correlations when squared give the percentage of

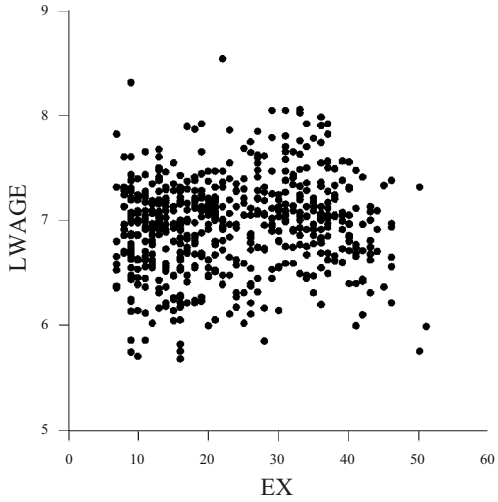


Figure 2.12 Log (Wage) versus Experience

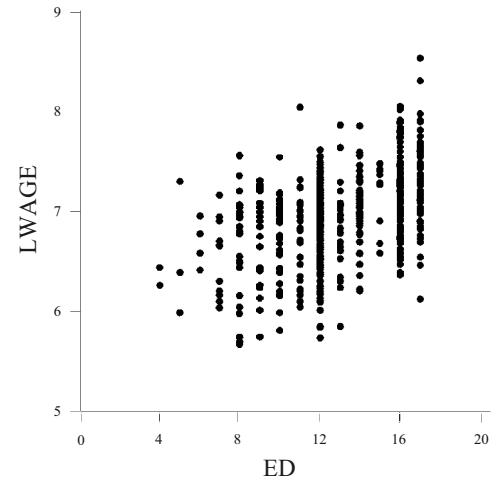


Figure 2.13 Log (Wage) versus Education

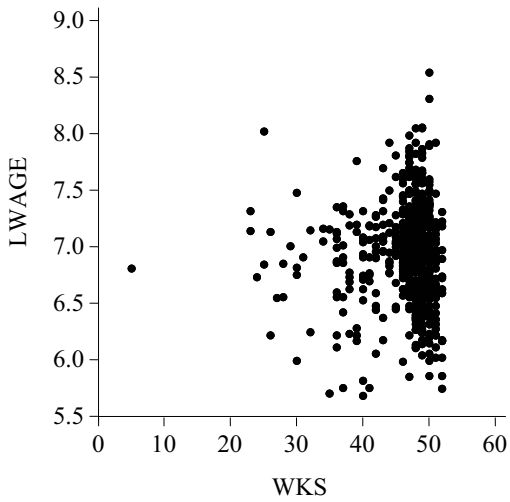


Figure 2.14 Log (Wage) versus Weeks

variation that one of these variables explain in the other. For example, the simple correlation coefficient between log wage and marital status is 0.32. This means that marital status explains $(0.32)^2$ or 10% of the variation in log wage.

One cannot emphasize enough how important it is to check one's data. It is important to compute the descriptive statistics, simple plots of the data and simple correlations. A wrong minimum or maximum could indicate some possible data entry errors. Troughs or peaks in these plots may indicate important events for time series data, like wars or recessions, or influential observations. More on this in Chapter 8. Simple correlation coefficients that equal one indicate perfectly collinear variables and warn of the failure of a linear regression that has both variables included among the regressors, see Chapter 4.

Notes

1. Actually $E(s^2) = \sigma^2$ does not need the normality assumption. This fact along with the proof of $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$, under Normality, can be easily shown using matrix algebra and is deferred to Chapter 7.
2. This can be proven using the Chebyshev's inequality, see Hogg and Craig (1995).
3. See Hogg and Craig (1995) for the type of regularity conditions needed for these distributions.

Problems

1. *Variance and Covariance of Linear Combinations of Random Variables.* Let a, b, c, d, e and f be arbitrary constants and let X and Y be two random variables.
   (a) Show that $\text{var}(a + bX) = b^2 \text{ var}(X)$.
   (b) $\text{var}(a + bX + cY) = b^2\text{var}(X) + c^2 \text{ var}(Y) + 2bc \text{ cov}(X, Y)$.
   (c) $\text{cov}[(a + bX + cY), (d + eX + fY)] = be \text{ var}(X) + cf \text{ var}(Y) + (bf + ce) \text{ cov}(X, Y)$.
2. *Independence and Simple Correlation.*
   (a) Show that if X and Y are independent, then $E(XY) = E(X)E(Y) = \mu_x\mu_y$ where $\mu_x = E(X)$ and $\mu_y = E(Y)$. Therefore, $\text{cov}(X, Y) = E(X - \mu_x)(Y - \mu_y) = 0$.
   (b) Show that if $Y = a + bX$, where a and b are arbitrary constants, then $\rho_{xy} = 1$ if $b > 0$ and -1 if $b < 0$.
3. *Zero Covariance Does Not Necessarily Imply Independence.* Let $X = -2, -1, 0, 1, 2$ with $\Pr[X = x] = 1/5$. Assume a perfect quadratic relationship between Y and X, namely $Y = X^2$. Show that $\text{cov}(X, Y) = E(X^3) = 0$. Deduce that $\rho_{XY} = \text{correlation } (X, Y) = 0$. The simple correlation coefficient ρ_{XY} measures the strength of the *linear* relationship between X and Y. For this example, it is zero even though there is a perfect *nonlinear* relationship between X and Y. This is also an example of the fact that if $\rho_{XY} = 0$, then X and Y are not necessarily independent. $\rho_{xy} = 0$ is a necessary but not sufficient condition for X and Y to be independent. The converse, however, is true, i.e., if X and Y are independent, then $\rho_{XY} = 0$, see problem 2.
4. The *Binomial Distribution* is defined as the number of successes in n independent Bernoulli trials with probability of success θ. This discrete probability function is given by

$$f(X;\theta) = \binom{n}{X}\theta^X(1-\theta)^{n-X} \quad X = 0, 1, \ldots, n$$

   and zero elsewhere, with $\binom{n}{X} = n!/[X!(n-X)!]$.
   (a) Out of 20 candidates for a job with a probability of hiring of 0.1. Compute the probabilities of getting $X = 5$ or 6 people hired?
   (b) Show that $\binom{n}{X} = \binom{n}{n-X}$ and use that to conclude that $\text{b}(n, X,) = \text{b}(n, n - X, 1 - \theta)$.
   (c) Verify that $E(X) = n\theta$ and $\text{var}(X) = n\theta(1 - \theta)$.
   (d) For a random sample of size n drawn from the Bernoulli distribution with parameter θ, show that $\bar{X}$ is the MLE of θ.
   (e) Show that $\bar{X}$, in part (d), is unbiased and consistent for θ.
   (f) Show that $\bar{X}$, in part (d), is sufficient for θ.

(g) Derive the Cramér-Rao lower bound for any unbiased estimator of θ. Is $\bar{X}$, in part (d), MVU for θ?

(h) For $n = 20$, derive the uniformly most powerful critical region of size $\alpha \leq 0.05$ for testing $H_0; \theta = 0.2$ versus $H_1; \theta = 0.6$. What is the probability of type II error for this test criteria?

(i) Form the Likelihood Ratio test for testing $H_0; \theta = 0.2$ versus $H_1; \theta \neq 0.2$. Derive the Wald and LM test statistics for testing H_0 versus H_1. When is the Wald statistic greater than the LM statistic?

5. For a random sample of size n drawn from the *Normal distribution* with mean μ and variance σ^2.

(a) Show that s^2 is a sufficient statistic for σ^2.

(b) Using the fact that $(n-1)s^2/\sigma^2$ is χ^2_{n-1} (without proof), verify that $E(s^2) = \sigma^2$ and that $\text{var}(s^2) = 2\sigma^4/(n-1)$ as shown in the text.

(c) Given that σ^2 is unknown, form the Likelihood Ratio test statistic for testing $H_0; \mu = 2$ versus $H_1; \mu \neq 2$. Derive the Wald and Lagrange Multiplier statistics for testing H_0 versus H_1. Verify that they are given by the expressions in example 4.

(d) Another derivation of the $W \geq LR \geq LM$ inequality for the null hypothesis given in part (c) can be obtained as follows: Let $\tilde{\mu}$, $\tilde{\sigma}^2$ be the restricted maximum likelihood estimators under $H_0; \mu = \mu_0$. Let $\hat{\mu}$, $\hat{\sigma}^2$ be the corresponding unrestricted maximum likelihood estimators under the alternative $H_1; \mu \neq \mu_0$. Show that $W = -2\log[L(\tilde{\mu}, \hat{\sigma}^2)/L(\hat{\mu}, \hat{\sigma}^2)]$ and $LM = -2\log[L(\tilde{\mu}, \tilde{\sigma}^2)/L(\hat{\mu}, \tilde{\sigma}^2)]$ where $L(\mu, \sigma^2)$ denotes the likelihood function. Conclude that $W \geq LR \geq LM$, see Breusch (1979). This is based on Baltagi (1994).

(e) Given that μ is unknown, form the Likelihood Ratio test statistic for testing $H_0; \sigma = 3$ versus $H_1; \sigma \neq 3$.

(f) Form the Likelihood Ratio test statistic for testing $H_0; \mu = 2, \sigma^2 = 4$ against the alternative that $H_1; \mu \neq 2$ or $\sigma^2 \neq 4$.

(g) For $n = 20, s^2 = 9$ construct a 95% confidence interval for σ^2.

6. The *Poisson* distribution can be defined as the limit of a Binomial distribution as $n \to \infty$ and $\theta \to 0$ such that $n\theta = \lambda$ is a positive constant. For example, this could be the probability of a rare disease and we are random sampling a large number of inhabitants, or it could be the rare probability of finding oil and n is the large number of drilling sights. This discrete probability function is given by

$$f(X; \lambda) = \frac{e^{-\lambda}\lambda^X}{X!} \quad X = 0, 1, 2, \ldots$$

For a random sample from this Poisson distribution

(a) Show that $E(X) = \lambda$ and $\text{var}(X) = \lambda$.

(b) Show that the MLE of λ is $\hat{\lambda}_{MLE} = \bar{X}$.

(c) Show that the method of moments estimator of λ is also $\bar{X}$.

(d) Show that $\bar{X}$ is unbiased and consistent for λ.

(e) Show that $\bar{X}$ is sufficient for λ.

(f) Derive the Cramér-Rao lower bound for any unbiased estimator of λ. Show that $\bar{X}$ attains that bound.

(g) For $n = 9$, derive the Uniformly Most Powerful critical region of size $\alpha \leq 0.05$ for testing $H_0; \lambda = 2$ versus $H_1; \lambda = 4$.

(h) Form the Likelihood Ratio test for testing H_0; $\lambda = 2$ versus H_1; $\lambda \neq 2$. Derive the Wald and LM statistics for testing H_0 versus H_1. When is the Wald test statistic greater than the LM statistic?

7. The *Geometric* distribution is known as the probability of waiting for the first success in independent repeated trials of a Bernoulli process. This could occur on the 1st, 2nd, 3rd,.. trials.

$$g(X;\theta) = \theta(1-\theta)^{X-1} \text{ for } X = 1, 2, 3, \ldots$$

(a) Show that $E(X) = 1/\theta$ and $\text{var}(X) = (1-\theta)/\theta^2$.

(b) Given a random sample from this Geometric distribution of size n, find the MLE of θ and the method of moments estimator of θ.

(c) Show that $\bar{X}$ is unbiased and consistent for $1/\theta$.

(d) For $n = 20$, derive the Uniformly Most Powerful critical region of size $\alpha \leq 0.05$ for testing H_0; $\theta = 0.5$ versus H_1; $\theta = 0.3$.

(e) Form the Likelihood Ratio test for testing H_0; $\theta = 0.5$ versus H_1; $\theta \neq 0.5$. Derive the Wald and LM statistics for testing H_0 versus H_1. When is the Wald statistic greater than the LM statistic?

8. The *Uniform* density, defined over the unit interval $[0, 1]$, assigns a unit probability for all values of X in that interval. It is like a roulette wheel that has an equal chance of stopping anywhere between 0 and 1.

$$\begin{aligned} f(X) &= 1 \qquad 0 \leq X \leq 1 \\ &= 0 \qquad \text{elsewhere} \end{aligned}$$

Computers are equipped with a Uniform (0,1) random number generator so it is important to understand these distributions.

(a) Show that $E(X) = 1/2$ and $\text{var}(X) = 1/12$.

(b) What is the $\Pr[0.1 < X < 0.3]$? Does it matter if we ask for the $\Pr[0.1 \leq X \leq 0.3]$?

9. The *Exponential* distribution is given by

$$f(X;\theta) = \frac{1}{\theta} e^{-X/\theta} \quad X > 0 \text{ and } \theta > 0$$

This is a skewed and continuous distribution defined only over the positive quadrant.

(a) Show that $E(X) = \theta$ and $\text{var}(X) = \theta^2$.

(b) Show that $\widehat{\theta}_{MLE} = \bar{X}$.

(c) Show that the method of moments estimator of θ is also $\bar{X}$.

(d) Show that $\bar{X}$ is an unbiased and consistent estimator of θ.

(e) Show that $\bar{X}$ is sufficient for θ.

(f) Derive the Cramér-Rao lower bound for any unbiased estimator of θ? Is $\bar{X}$ MVU for θ?

(g) For $n = 20$, derive the Uniformly Most Powerful critical region of size $\alpha \leq 0.05$ for testing $H_0; \theta = 1$ versus $H_1; \theta = 2$.

(h) Form the Likelihood Ratio test for testing $H_0; \theta = 1$ versus $H_1; \theta \neq 1$. Derive the Wald and LM statistics for testing H_0 versus H_1. When is the Wald statistic greater than the LM statistic?

10. The *Gamma* distribution is given by

$$\begin{aligned} f(X;\alpha,\beta) &= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} X^{\alpha-1} e^{-X/\beta} && \text{for } X > 0 \\ &= 0 && \text{elsewhere} \end{aligned}$$

where α and $\beta > 0$ and $\Gamma(\alpha) = (\alpha - 1)!$ This is a skewed and continuous distribution.

(a) Show that $E(X) = \alpha\beta$ and $\text{var}(X) = \alpha\beta^2$.

(b) For a random sample drawn from this Gamma density, what are the method of moments estimators of α and β?

(c) Verify that for $\alpha = 1$ and $\beta = \theta$, the Gamma probability density function reverts to the Exponential p.d.f. considered in problem 9.

(d) We state without proof that for $\alpha = r/2$ and $\beta = 2$, this Gamma density reduces to a χ^2 distribution with r degrees of freedom, denoted by χ^2_r. Show that $E(\chi^2_r) = r$ and $\text{var}(\chi^2_r) = 2r$.

(e) For a random sample from the χ^2_r distribution, show that $(X_1 X_2, \ldots, X_n)$ is a sufficient statistic for r.

(f) One can show that the square of a $N(0,1)$ random variable is a χ^2 random variable with 1 degree of freedom, see the Appendix to the chapter. Also, one can show that the sum of independent χ^2's is a χ^2 random variable with degrees of freedom equal the sum of the corresponding degrees of freedom of the individual χ^2's, see problem 15. This will prove useful for testing later on. Using these results, verify that the sum of squares of m independent $N(0,1)$ random variables is a χ^2 with m degrees of freedom.

11. The *Beta* distribution is defined by

$$\begin{aligned} f(X) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} X^{\alpha-1}(1-X)^{\beta-1} && \text{for } 0 < X < 1 \\ &= 0 && \text{elsewhere} \end{aligned}$$

where $\alpha > 0$ and $\beta > 0$. This is a skewed continuous distribution.

(a) For $\alpha = \beta = 1$ this reverts back to the Uniform $(0,1)$ probability density function. Show that $E(X) = (\alpha/\alpha + \beta)$ and $\text{var}(X) = \alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1)$.

(b) Suppose that $\alpha = 1$, find the estimators of β using the method of moments and the method of maximum likelihood.

12. The *t-distribution* with r degrees of freedom can be defined as the ratio of two independent random variables. The numerator being a $N(0,1)$ random variable and the denominator being the square-root of a χ^2_r random variable divided by its degrees of freedom. The t-distribution is a symmetric distribution like the Normal distribution but with fatter tails. As $r \to \infty$, the t-distribution approaches the Normal distribution.

(a) Verify that if $X_1, \ldots, X_n$ are a random sample drawn from a $N(\mu, \sigma^2)$ distribution, then $z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ is $N(0,1)$.

(b) Use the fact that $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ to show that $t = z/\sqrt{s^2/\sigma^2} = (\bar{X} - \mu)/(s/\sqrt{n})$ has a t-distribution with $(n-1)$ degrees of freedom. We use the fact that s^2 is independent of $\bar{X}$ without proving it.

(c) For $n = 16$, $\bar{x} = 20$ and $s^2 = 4$, construct a 95% confidence interval for μ.

13. The *F-distribution* can be defined as the ratio of two independent χ^2 random variables each divided by its corresponding degrees of freedom. It is commonly used to test the equality of variances. Let s_1^2 be the sample variance from a random sample of size n_1 drawn from $N(\mu_1, \sigma_1^2)$ and let s_2^2 be the sample variance from another random sample of size n_2 drawn from $N(\mu_2, \sigma_2^2)$. We know that $(n_1-1)s_1^2/\sigma_1^2$ is $\chi^2_{(n_1-1)}$ and $(n_2-1)s_2^2/\sigma_2^2$ is $\chi^2_{(n_2-1)}$. Taking the ratio of those two independent χ^2 random variables divided by their appropriate degrees of freedom yields

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

which under the null hypothesis $H_0; \sigma_1^2 = \sigma_2^2$ gives $F = s_1^2/s_2^2$ and is distributed as F with (n_1-1) and (n_2-1) degrees of freedom. Both s_1^2 and s_2^2 are observable, so F can be computed and compared to critical values for the F-distribution with the appropriate degrees of freedom. Two inspectors drawing two random samples of size 25 and 31 from two shifts of a factory producing steel rods, find that the sampling variance of the lengths of these rods are 15.6 and 18.9 inches squared. Test whether the variances of the two shifts are the same.

14. *Moment Generating Function (MGF).*

    (a) Derive the MGF of the Binomial distribution defined in problem 4. Show that it is equal to $[(1-\theta)+\theta e^t]^n$.

    (b) Derive the MGF of the Normal distribution defined in problem 5. Show that it is $e^{\mu t + \frac{1}{2}\sigma^2 t^2}$.

    (c) Derive the MGF of the Poisson distribution defined in problem 6. Show that it is $e^{\lambda(e^t-1)}$.

    (d) Derive the MGF of the Geometric distribution defined in problem 7. Show that it is $\theta e^t/[1-(1-\theta)e^t]$.

    (e) Derive the MGF of the Exponential distribution defined in problem 9. Show that it is $1/(1-\theta t)$.

    (f) Derive the MGF of the Gamma distribution defined in problem 10. Show that it is $(1-\beta t)^{-\alpha}$. Conclude that the MGF of a χ_r^2 is $(1-2t)^{-\frac{r}{2}}$.

    (g) Obtain the mean and variance of each distribution by differentiating the corresponding MGF derived in parts (a) through (f).

15. *Moment Generating Function Method.*

    (a) Show that if $X_1, \ldots, X_n$ are independent Poisson distributed with parameters (λ_i) respectively, then $Y = \sum_{i=1}^n X_i$ is Poisson with parameter $\sum_{i=1}^n \lambda_i$.

    (b) Show that if $X_1, \ldots, X_n$ are independent Normally distributed with parameters (μ_i, σ_i^2), then $Y = \sum_{i=1}^n X_i$ is Normal with mean $\sum_{i=1}^n \mu_i$ and variance $\sum_{i=1}^n \sigma_i^2$.

    (c) Deduce from part (b) that if $X_1, \ldots, X_n$ are IIN(μ, σ^2), then $\bar{X} \sim N(\mu, \sigma^2/n)$.

    (d) Show that if $X_1, \ldots, X_n$ are independent χ^2 distributed with parameters (r_i) respectively, then $Y = \sum_{i=1}^n X_i$ is χ^2 distributed with parameter $\sum_{i=1}^n r_i$.

16. *Best Linear Prediction.* (Problems 16 and 17 are based on Amemiya (1994)). Let X and Y be two random variables with means μ_X and μ_Y and variances σ_X^2 and σ_Y^2, respectively. Suppose that

$$\rho = \text{correlation}(X, Y) = \sigma_{XY}/\sigma_X\sigma_Y$$

where $\sigma_{XY} = \text{cov}(X, Y)$. Consider the *linear* relationship $Y = \alpha + \beta X$ where α and β are scalars:

    (a) Show that the *best linear predictor* of Y based on X, where *best* in this case means the minimum mean squared error predictor which minimizes $E(Y-\alpha-\beta X)^2$ with respect to α and β is given by $\widehat{Y} = \widehat{\alpha} + \widehat{\beta} X$ where $\widehat{\alpha} = \mu_Y - \widehat{\beta}\mu_X$ and $\widehat{\beta} = \sigma_{XY}/\sigma_X^2 = \rho\sigma_Y/\sigma_X$.

(b) Show that the $\text{var}(\widehat{Y}) = \rho^2\sigma_Y^2$ and that $\widehat{u} = Y - \widehat{Y}$, the prediction error, has mean zero and variance equal to $(1-\rho^2)\sigma_Y^2$. Therefore, ρ^2 can be interpreted as the proportion of σ_Y^2 that is explained by the *best linear* predictor $\widehat{Y}$.

(c) Show that $\text{cov}(\widehat{Y}, \widehat{u}) = 0$.

17. *The Best Predictor.* Let X and Y be the two random variables considered in problem 16. Now consider predicting Y by a general, possibly non-linear, function of X denoted by $h(X)$.

(a) Show that the *best predictor* of Y based on X, where *best* in this case means the minimum mean squared error predictor that minimizes $E[Y - h(X)]^2$ is given by $h(X) = E(Y/X)$. **Hint:** Write $E[Y - h(X)]^2$ as $E\{[Y - E(Y/X)] + [E(Y/X) - h(X)]\}^2$. Expand the square and show that the cross-product term has zero expectation. Conclude that this mean squared error is minimized at $h(X) = E(Y/X)$.

(b) If X and Y are bivariate Normal, show that the *best predictor* of Y based on X is identical to the *best linear predictor* of Y based on X.

18. *Descriptive Statistics.* Using the data used in section 2.6 based on 595 individuals drawn from the Panel Study of Income Dynamics for 1982 and available on the Springer web site as EARN.ASC, replicate the tables and graphs given in that section. More specifically

(a) replicate Table 2.1 which gives the descriptive statistics for a subset of the variables in this data set.

(b) Replicate Figures 2.6–2.11 which plot the histograms for log wage, weeks worked, education and experience.

(c) Replicate Table 2.2 which gives the average log wage for various groups and test the difference between these averages using a t-test.

(d) Replicate Figure 2.12 which plots log wage versus experience. Figure 2.13 which plots log wage versus education and Figure 2.14 which plots log wage versus weeks worked.

(e) Replicate Table 2.3 which gives the correlation matrix among a subset of these variables.

19. *Conflict Among Criteria for Testing Hypotheses: Examples from Non-Normal Distributions.* This is based on Baltagi (2000). Berndt and Savin (1977) showed that $W \geq LR \geq LM$ for the case of a multivariate regression model with normal distrubances. Ullah and Zinde-Walsh (1984) showed that this inequality is not robust to non-normality of the disturbances. In the spirit of the latter article, this problem considers simple examples from non-normal distributions and illustrates how this conflict among criteria is affected.

(a) Consider a random sample $x_1, x_2, \ldots, x_n$ from a Poisson distribution with parameter λ. Show that for testing $\lambda = 3$ versus $\lambda \neq 3$ yields $W \geq LM$ for $\bar{x} \leq 3$ and $W \leq LM$ for $\bar{x} \geq 3$.

(b) Consider a random sample $x_1, x_2, \ldots, x_n$ from an Exponential distribution with parameter θ. Show that for testing $\theta = 3$ versus $\theta \neq 3$ yields $W \geq LM$ for $0 < \bar{x} \leq 3$ and $W \leq LM$ for $\bar{x} \geq 3$.

(c) Consider a random sample $x_1, x_2, \ldots, x_n$ from a Bernoulli distribution with parameter θ. Show that for testing $\theta = 0.5$ versus $\theta \neq 0.5$, we will always get $W \geq LM$. Show also, that for testing $\theta = (2/3)$ versus $\theta \neq (2/3)$ we get $W \leq LM$ for $(1/3) \leq \bar{x} \leq (2/3)$ and $W \geq LM$ for $(2/3) \leq \bar{x} \leq 1$ or $0 < \bar{x} \leq (1/3)$.

References

More detailed treatment of the material in this chapter may be found in:

Amemiya, T. (1994), *Introduction to Statistics and Econometrics* (Harvard University Press: Cambridge).

Baltagi, B.H. (1994), "The Wald, LR, and LM Inequality," *Econometric Theory*, Problem 94.1.2, 10: 223-224.

Baltagi, B.H. (2000), "Conflict Among Critera for Testing Hypotheses: Examples from Non-Normal Distributions," *Econometric Theory*, Problem 00.2.4, 16: 288.

Bera A.K. and G. Permaratne (2001), "General Hypothesis Testing," Chapter 2 in Baltagi, B.H. (ed.), *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Berndt, E.R. and N.E. Savin (1977), "Conflict Among Criteria for Testing Hypotheses in the Multivariate Linear Regression Model," *Econometrica*, 45: 1263-1278.

Breusch, T.S. (1979), "Conflict Among Criteria for Testing Hypotheses: Extensions and Comments," *Econometrica*, 47: 203-207.

Buse, A. (1982), "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note," *The American Statistician,* 36 :153-157.

DeGroot, M.H. (1986), *Probability and Statistics* (Addison-Wesley: Mass.).

Freedman, D., R. Pisani, R. Purves and A. Adhikari (1991), *Statistics* (Norton: New York).

Freund, J.E. (1992), *Mathematical Statistics* (Prentice-Hall: New Jersey).

Hogg, R.V. and A.T. Craig (1995), *Introduction to Mathematical Statistics* (Prentice Hall: New Jersey).

Jollife, I.T. (1995), "Sample Sizes and the Central Limit Theorem: The Poisson Distribution as an Illustration," *The American Statistician*, 49: 269.

Kennedy, P. (1992), *A Guide to Econometrics* (MIT Press: Cambridge).

Mood, A.M., F.A. Graybill and D.C. Boes (1974), *Introduction to the Theory of Statistics* (McGraw-Hill: New York).

Spanos, A. (1986), *Statistical Foundations of Econometric Modelling* (Cambridge University Press: Cambridge).

Ullah, A. and V. Zinde-Walsh (1984), "On the Robustness of LM, LR and W Tests in Regression Models," *Econometrica*, 52: 1055-1065.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics* (Wiley: New York).

Appendix

Score and Information Matrix: The likelihood function of a sample $X_1, \ldots, X_n$ drawn from $f(X_i, \theta)$ is really the joint probability density function written as a function of θ:

$$L(\theta) = f(X_1, \ldots, X_n; \theta)$$

This probability density function has the property that $\int L(\theta) d\mathbf{x} = 1$ where the integral is over all $X_1, \ldots, X_n$ written compactly as one integral over $\mathbf{x}$. Differentiating this multiple integral with respect

to θ, one gets

$$\int \frac{\partial L}{\partial \theta} d\mathbf{x} = 0$$

Multiplying and dividing by L, one gets

$$\int \left(\frac{1}{L}\frac{\partial L}{\partial \theta}\right) L \, d\mathbf{x} = \int \left(\frac{\partial \log L}{\partial \theta}\right) L \, d\mathbf{x} = 0$$

But the *score* is by definition $S(\theta) = \partial \log L/\partial\theta$. Hence $E[S(\theta)] = 0$. Differentiating again with respect to θ, one gets

$$\int \left[\left(\frac{\partial^2 \log L}{\partial \theta^2}\right) L + \int \left(\frac{\partial \log L}{\partial \theta}\right)\left(\frac{\partial L}{\partial \theta}\right)\right] d\mathbf{x} = 0$$

Multiplying and dividing the second term by L one gets

$$E\left[\frac{\partial^2 \log L}{\partial \theta^2} + \left(\frac{\partial \log L}{\partial \theta}\right)^2\right] = 0$$

or

$$E\left[-\frac{\partial^2 \log L}{\partial \theta^2}\right] = E\left[\left(\frac{\partial \log L}{\partial \theta}\right)^2\right] = E[S(\theta)]^2$$

But $\text{var}[S(\theta)] = E[S(\theta)]^2$ since $E[S(\theta)] = 0$. Hence $\text{I}(\theta) = \text{var}[S(\theta)]$.

Moment Generating Function (MGF): For the random variable X, the expected value of a special function of X, namely e^{Xt} is denoted by

$$M_X(t) = E(e^{Xt}) = E(1 + Xt + X^2\frac{t^2}{2!} + X^3\frac{t^3}{3!} + ..)$$

where the second equality follows from the Taylor series expansion of e^{Xt} around zero. Therefore,

$$M_X(t) = 1 + E(X)t + E(X^2)\frac{t^2}{2!} + E(X^3)\frac{t^3}{3!} + ..$$

This function of t generates the moments of X as coefficients of an infinite polynomial in t. For example, $\mu = E(X) =$ coefficient of t, and $E(X^2)/2$ is the coefficient of t^2, etc. Alternatively, one can differentiate this MGF with respect to t and obtain $\mu = E(X) = M'_X(0)$, i.e., the first derivative of $M_X(t)$ with respect to t evaluated at $t = 0$. Similarly, $E(X^r) = M^r_X(0)$ which is the r-th derivative of $M_X(t)$ with respect to t evaluated at $t = 0$. For example, for the Bernoulli distribution;

$$M_X(t) = E(e^{Xt}) = \textstyle\sum_{X=0}^{1} e^{Xt}\theta^X(1-\theta)^{1-X} = \theta e^t + (1-\theta)$$

so that $M'_X(t) = \theta e^t$ and $M'_X(0) = \theta = E(X)$ and $M''_X(t) = \theta e^t$ which means that $E(X^2) = M''_X(0) = \theta$.

Hence,

$$\text{var}(X) = E(X^2) - (E(X))^2 = \theta - \theta^2 = \theta(1-\theta).$$

For the Normal distribution, see problem 14, it is easy to show that if $X \sim N(\mu, \sigma^2)$, then $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$ and $M'_X(0) = E(X) = \mu$ and $M''_X(0) = E(X^2) = \sigma^2 + \mu^2$.

There is a one-to-one correspondence between the MGF when it exists and the corresponding p.d.f. This means that if Y has a MGF given by e^{2t+4t^2} then Y is normally distributed with mean 2 and variance 8. Similarly, if Z has a MGF given by $(e^t + 1)/2$, then Z is Bernoulli distributed with $\theta = 1/2$.

Change of Variable: If $X \sim N(0,1)$, then one can find the distribution function of $Y = |X|$ by using the *Distribution Function* method. By definition the distribution function of y is defined as

$$\begin{aligned} G(y) &= \Pr[Y \le y] = \Pr[|X| \le y] = \Pr[-y \le X \le y] \\ &= \Pr[X \le y] - \Pr[X \le -y] = F(y) - F(-y) \end{aligned}$$

so that the distribution function of $Y, G(y)$, can be obtained from the distribution function of X, $F(x)$. Since the $N(0,1)$ distribution is symmetric around zero, then $F(-y) = 1 - F(y)$ and substituting that in $G(y)$ we get $G(y) = 2F(y) - 1$. Recall, that the p.d.f. of Y is given by $g(y) = G'(y)$. Hence, $g(y) = f(y) + f(-y)$ and this reduces to $2f(y)$ if the distribution is symmetric around zero. So that if $f(x) = e^{-x^2/2}/\sqrt{2\pi}$ for $-\infty < x < +\infty$ then $g(y) = 2f(y) = 2e^{-y^2/2}/\sqrt{2\pi}$ for $y \ge 0$.

Let us now find the distribution of $Z = X^2$, the square of a $N(0,1)$ random variable. Note that $dZ/dX = 2X$ which is positive when $X > 0$ and negative when $X < 0$. The *change of variable* method cannot be applied since $Z = X^2$ is not a monotonic transformation over the entire domain of X. However, using $Y = |X|$, we get $Z = Y^2 = (|X|)^2$ and $dZ/dY = 2Y$ which is always non-negative since Y is non-negative. In this case, the *change of variable* method states that the p.d.f. of Z is obtained from that of Y by substituting the inverse transformation $Y = \sqrt{Z}$ into the p.d.f. of Y and multiplying it by the absolute value of the derivative of the inverse transformation:

$$h(z) = g(\sqrt{z}) \cdot |\frac{dY}{dZ}| = \frac{2}{\sqrt{2\pi}} e^{-z/2} |\frac{1}{2\sqrt{z}}| = \frac{1}{\sqrt{2\pi}} z^{-1/2} e^{-z/2} \text{ for } z \ge 0$$

It is clear why this transformation will not work for X since $Z = X^2$ has two solutions for the inverse transformation, $X = \pm\sqrt{Z}$, whereas, there is one unique solution for $Y = \sqrt{Z}$ since it is non-negative. Using the results of problem 10, one can deduce that Z has a gamma distribution with $\alpha = 1/2$ and $\beta = 2$. This special Gamma density function is a χ^2 distribution with 1 degree of freedom. Hence, we have shown that the square of a $N(0,1)$ random variable has a χ^2_1 distribution.

Finally, if $X_1, \ldots, X_n$ are independently distributed then the distribution function of $Y = \sum_{i=1}^n X_i$ can be obtained from that of the X_i's using the *Moment Generating Function* (MGF) method:

$$\begin{aligned} M_Y(t) &= E(e^{Yt}) = E[e^{(\sum_{i=1}^n X_i)t}] = E(e^{X_1 t})E(e^{X_2 t})..E(e^{X_n t}) \\ &= M_{X_1}(t)M_{X_2}(t)..M_{X_n}(t) \end{aligned}$$

If in addition these X_i's are identically distributed, then $M_{X_i}(t) = M_X(t)$ for $i = 1, \ldots, n$ and

$$M_Y(t) = [M_X(t)]^n$$

For example, if $X_1, \ldots, X_n$ are IID Bernoulli (θ), then $M_{X_i}(t) = M_X(t) = \theta e^t + (1-\theta)$ for $i = 1, \ldots, n$. Hence the MGF of $Y = \sum_{i=1}^n X_i$ is given by

$$M_Y(t) = [M_X(t)]^n = [\theta e^t + (1-\theta)]^n$$

This can be easily shown to be the MGF of the Binomial distribution given in problem 14. This proves that the sum of n independent and identically distributed Bernoulli random variables with parameter θ is a Binomial random variable with same parameter θ.

Central Limit Theorem: If $X_1, \ldots, X_n$ are IID(μ, σ^2) from an unknown distribution, then $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is asymptotically distributed as $N(0,1)$.

Proof: We assume that the MGF of the X_i's exist and derive the MGF of Z. Next, we show that lim $M_Z(t)$ as $n \to \infty$ is $e^{1/2t^2}$ which is the MGF of $N(0,1)$ distribution. First, note that

$$Z = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} = \frac{Y - n\mu}{\sigma\sqrt{n}}$$

where $Y = \sum_{i=1}^{n} X_i$ with $M_Y(t) = [M_X(t)]^n$. Therefore,

$$\begin{aligned} M_Z(t) &= E(e^{Zt}) = E\left(e^{(Yt-n\mu t)/\sigma\sqrt{n}}\right) = e^{-n\mu t/\sigma\sqrt{n}} E\left(e^{Yt/\sigma\sqrt{n}}\right) \\ &= e^{-n\mu t/\sigma\sqrt{n}} M_Y(t/\sigma\sqrt{n}) = e^{-n\mu t/\sigma\sqrt{n}} [M_X(t/\sigma\sqrt{n})]^n \end{aligned}$$

Taking log of both sides we get

$$\log M_Z(t) = \frac{-n\mu t}{\sigma\sqrt{n}} + n\log[1 + \frac{t}{\sigma\sqrt{n}}E(X) + \frac{t^2}{2\sigma^2 n}E(X^2) + \frac{t^3}{6\sigma^3 n\sqrt{n}}E(X^3) + ..]$$

Using the Taylor series expansion $\log(1+s) = s - \frac{s^2}{2} + \frac{s^3}{3} - ..$ we get

$$\begin{aligned} \log M_Z(t) &= -\frac{\sqrt{n}\mu}{\sigma}t + n\left\{\left[\mu\frac{t}{\sigma\sqrt{n}} + \frac{t^2}{2\sigma^2 n}E(X^2) + \frac{t^3}{6\sigma^3 n\sqrt{n}}E(X^3) + ..\right]\right. \\ &\quad -\frac{1}{2}\left[\mu\frac{t}{\sigma\sqrt{n}} + \frac{t^2}{2\sigma^2 n}E(X^2) + \frac{t^3}{6\sigma^3 n\sqrt{n}}E(X^3) + ..\right]^2 \\ &\quad \left.+\frac{1}{3}\left[\mu\frac{t}{\sigma\sqrt{n}} + \frac{t^2}{2\sigma^2\sqrt{n}}E(X^2) + \frac{t^3}{6\sigma^3 n\sqrt{n}}E(X^3) + ..\right]^3 - ..\right\} \end{aligned}$$

Collecting powers of t, we get

$$\begin{aligned} \log M_Z(t) &= \left(-\frac{\sqrt{n}\mu}{\sigma} + \frac{\sqrt{n}\mu}{\sigma}\right)t + \left(\frac{(E(X^2)}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right)t^2 \\ &\quad + \left(\frac{E(X^3)}{6\sigma^3\sqrt{n}} - \frac{1}{2}\cdot\frac{2\mu E(X^2)}{2\sigma^3\sqrt{n}} + \frac{1}{3}\frac{\mu^3}{\sigma^3\sqrt{n}}\right)t^3 + .. \end{aligned}$$

Therefore

$$\log M_Z(t) = \frac{1}{2}t^2 + \left(\frac{E(X^3)}{6} - \frac{\mu E(X^2)}{2} + \frac{\mu^3}{3}\right)\frac{t^3}{\sigma^3\sqrt{n}} + ..$$

note that the coefficient of t^3 is $1/\sqrt{n}$ times a constant. Therefore, this coefficient goes to zero as $n \to \infty$. Similarly, it can be shown that the coefficient of t^r is $1/\sqrt{n^{r-2}}$ times a constant for $r \geq 3$. Hence,

$$\lim_{n\to\infty} \log M_Z(t) = \frac{1}{2}t^2 \quad \text{and} \quad \lim_{n\to\infty} M_Z(t) = e^{\frac{1}{2}t^2}$$

which is the MGF of a standard normal distribution.

The Central Limit Theorem is a powerful tool for asymptotic inference. In real life we do not know what distribution we are sampling from, but as long as the sample drawn is random and we average (or sum) and standardize then as $n \to \infty$, the resulting standardized statistic has an asymptotic $N(0,1)$ distribution that can be used for inference.

Using a random number generator from say the uniform distribution on the computer, one can generate samples of size $n = 20, 30, 50$ from this distribution and show how the sampling distribution of the sum (or average) when it is standardized closely approximates the $N(0,1)$ distribution.

The real question for the applied researcher is how large n should be to invoke the Central Limit Theorem. This depends on the distribution we are drawing from. For a Bernoulli distribution, a larger n is needed the more asymmetric this distribution is i.e., if $\theta = 0.1$ rather than 0.5.

In fact, Figure 2.15 shows the Poisson distribution with mean = 15. This looks like a good approximation for a Normal distribution even though it is a discrete probability function. Problem 15 shows that the sum of n independent identically distributed Poisson random variables with parameter λ is a Poisson random variable with parameter $(n\lambda)$. This means that if $\lambda = 0.15$, an n of 100 will lead to the distribution of the sum being Poisson $(n\lambda = 15)$ and the Central Limit Theorem seems well approximated.

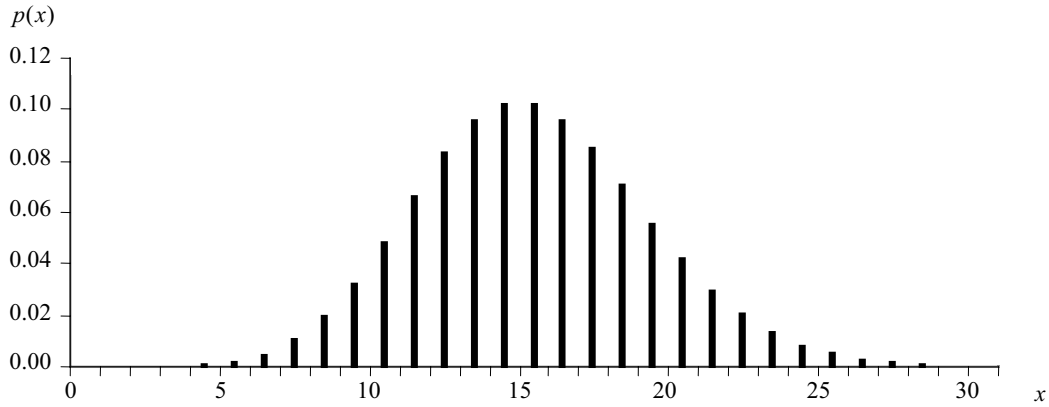


Figure 2.15 Poisson Probability Distribution, Mean = 15

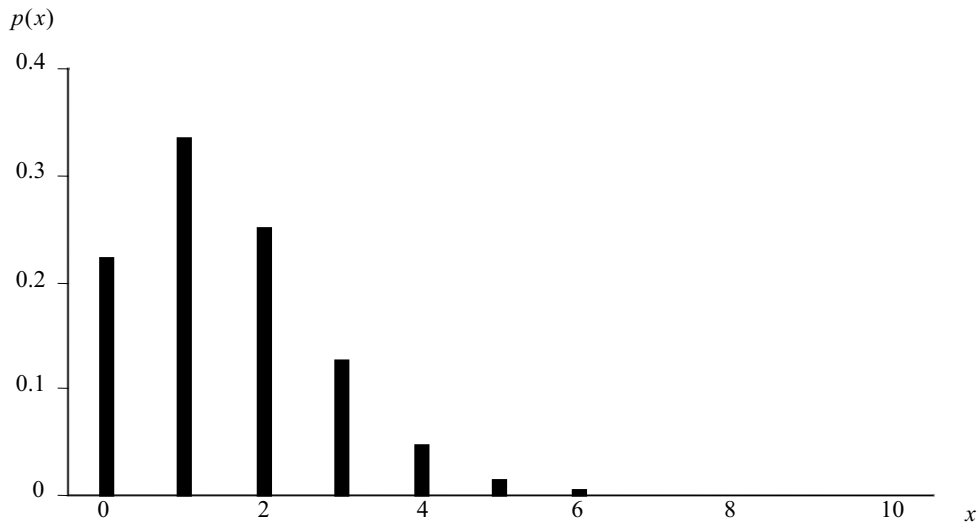


Figure 2.16 Poisson Probability Distribution, Mean = 1.5

However, if $\lambda = 0.015$, an n of 100 will lead to the distribution of the sum being Poisson ($n\lambda = 1.5$) which is given in Figure 2.16. This Poisson probability function is skewed and discrete and does not approximate well a normal density. This shows that one has to be careful in concluding that $n = 100$ is a large enough sample for the Central Limit Theorem to apply. We showed in this simple example that this depends on the distribution we are sampling from. This is true for Poisson ($\lambda = 0.15$) but not Poisson ($\lambda = 0.015$), see Joliffe (1995). The same idea can be illustrated with a skewed Bernoulli distribution.

Conditional Mean and Variance: Two random variables X and Y are bivariate Normal if they have the following joint distribution:

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)\right]\right\}$$

where $-\infty < x < +\infty$, $-\infty < y < +\infty$, $E(X) = \mu_X$, $E(Y) = \mu_Y$, $\text{var}(X) = \sigma_X^2$, $\text{var}(Y) = \sigma_Y^2$ and $\rho =$ correlation $(X, Y) = \text{cov}(X, Y)/\sigma_X \sigma_Y$. This joint density can be rewritten as

$$f(x,y) = \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left[y - \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]^2\right\}$$
$$\cdot\frac{1}{\sqrt{2\pi}\sigma_X} \exp\left\{-\frac{1}{2\sigma_X^2}(x-\mu_X)^2\right\} = f(y/x)f_1(x)$$

where $f_1(x)$ is the *marginal density* of X and $f(y/x)$ is the *conditional density* of Y given X. In this case, $X \sim N(\mu_X, \sigma_X^2)$ and Y/X is Normal with mean $E(Y/X) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_x)$ and variance given by $\text{var}(Y/X) = \sigma_Y^2(1-\rho^2)$.

By symmetry, the roles of X and Y can be interchanged and one can write $f(x,y) = f(x/y)\ f_2(y)$ where $f_2(y)$ is the marginal density of Y. In this case, $Y \sim N(\mu_Y, \sigma_Y^2)$ and X/Y is Normal with mean $E(X/Y) = \mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y - \mu_Y)$ and variance given by $\text{var}(X/Y) = \sigma_X^2(1-\rho^2)$. If $\rho = 0$, then $f(y/x) = f_2(y)$ and $f(x,y) = f_1(x)f_2(y)$ proving that X and Y are independent. Therefore, if $\text{cov}(X,Y) = 0$ *and* X and Y are bivariate Normal, then X and Y are independent. In general, $\text{cov}(X,Y) = 0$ alone does not necessarily imply independence, see problem 3.

One important and useful property is the *law of iterated expectations*. This says that the expectation of any function of X and Y say $h(X,Y)$ can be obtained as follows:

$$E[h(X,Y)] = E_X E_{Y/X}[h(X,Y)]$$

where the subscript Y/X on E means the conditional expectation of Y given that X is treated as a constant. The next expectation E_X treats X as a random variable. The proof is simple.

$$E[h(X,Y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} h(x,y)f(x,y)dxdy$$

where $f(x,y)$ is the joint density of X and Y. But $f(x,y)$ can be written as $f(y/x)f_1(x)$, hence $E[h(X,Y)] = \int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty} h(x,y)f(y/x)dy\right] f_1(x)dx = E_X E_{Y/X}[h(X,Y)]$.

Example: This law of iterated expectation can be used to show that for the bivariate Normal density, the parameter ρ is indeed the correlation coefficient of X and Y. In fact, let $h(X,Y) = XY$, then

$$\begin{aligned} E(XY) &= E_X E_{Y/X}(XY/X) = E_X X E(Y/X) = E_X X[\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X)] \\ &= \mu_X\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}\sigma_X^2 = \mu_X\mu_Y + \rho\sigma_Y\sigma_X \end{aligned}$$

Rearranging terms, one gets $\rho = [E(XY) - \mu_X\mu_Y]/\sigma_X\sigma_Y = \sigma_{XY}/\sigma_X\sigma_Y$ as required.

Another useful result pertains to the unconditional variance of $h(X,Y)$ being the sum of the mean of the conditional variance and the variance of the conditional mean:

$$\text{var}(h(X,Y)) = E_X\text{var}_{Y/X}[h(X,Y)] + \text{var}_X E_{Y/X}[h(X,Y)]$$

Proof: We will write $h(X,Y)$ as h to simplify the presentation

$$\text{var}_{Y/X}(h) = E_{Y/X}(h^2) - [E_{Y/X}(h)]^2$$

and taking expectations with respect to X yields $E_X\text{var}_{Y/X}(h) = E_X E_{Y/X}(h^2) - E_X[E_{Y/X}(h)]^2$
$= E(h^2) - E_X[E_{Y/X}(h)]^2$.

Also, $\text{var}_X E_{Y/X}(h) = E_X[E_{Y/X}(h)]^2 - (E_X[E_{Y/X}(h)])^2 = E_X[E_{Y/X}(h)]^2 - [E(h)]^2$ adding these two terms yields

$$E(h^2) - [E(h)]^2 = \text{var}(h).$$

CHAPTER 3
Simple Linear Regression

3.1 Introduction

In this chapter, we study extensively the estimation of a linear relationship between two variables, Y_i and X_i, of the form:

$$Y_i = \alpha + \beta X_i + u_i \quad i = 1, 2, \ldots, n \tag{3.1}$$

where Y_i denotes the i-th observation on the dependent variable Y which could be consumption, investment or output, and X_i denotes the i-th observation on the independent variable X which could be disposable income, the interest rate or an input. These observations could be collected on firms or households at a given point in time, in which case we call the data a cross-section. Alternatively, these observations may be collected over time for a specific industry or country in which case we call the data a time-series. n is the number of observations, which could be the number of firms or households in a cross-section, or the number of years if the observations are collected annually. α and β are the intercept and slope of this simple linear relationship between Y and X. They are assumed to be unknown parameters to be estimated from the data. A plot of the data, i.e., Y versus X would be very illustrative showing what type of relationship exists empirically between these two variables. For example, if Y is consumption and X is disposable income then we would expect a positive relationship between these variables and the data may look like Figure 3.1 when plotted for a random sample of households. If α and β were known, one could draw the straight line $(\alpha + \beta X)$ as shown in Figure 3.1. It is clear that not all the observations (X_i, Y_i) lie on the straight line $(\alpha + \beta X)$. In fact, equation (3.1) states that the difference between each Y_i and the corresponding $(\alpha + \beta X_i)$ is due to a random error u_i. This error may be due to (i) the omission of relevant factors that could influence consumption, other than disposable income, like real wealth or varying tastes, or unforseen events that induce households to consume more or less, (ii) measurement error, which could be the result of households not reporting their consumption or income accurately, or (iii) wrong choice of a linear relationship between consumption and income, when the true relationship may be nonlinear. These different causes of the error term will have different effects on the distribution of this error. In what follows, we consider only disturbances that satisfy some restrictive assumptions. In later chapters we relax these assumptions to account for more general kinds of error terms.

In real life, α and β are not known, and have to be estimated from the observed data $\{(X_i, Y_i)$ for $i = 1, 2, \ldots, n\}$. This also means that the true line $(\alpha + \beta X)$ as well as the true disturbances (the u_i's) are unobservable. In this case, α and β could be estimated by the best fitting line through the data. Different researchers may draw different lines through the same data. What makes one line better than another? One measure of misfit is the amount of error from the observed Y_i to the guessed line, let us call the latter $\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta} X_i$, where the hat ($\widehat{}$) denotes a guess on the appropriate parameter or variable. Each observation (X_i, Y_i) will have a corresponding observable error attached to it, which we will call $e_i = Y_i - \widehat{Y}_i$, see Figure 3.2. In other words, we obtain the guessed $Y_i, (\widehat{Y}_i)$ corresponding to each X_i from the guessed line,

$\widehat{\alpha} + \widehat{\beta}X_i$. Next, we find our error in guessing that Y_i, by subtracting the actual Y_i from the guessed $\widehat{Y}_i$. The only difference between Figure 3.1 and Figure 3.2 is the fact that Figure 3.1 draws the true consumption line which is unknown to the researcher, whereas Figure 3.2 is a guessed consumption line drawn through the data. Therefore, while the u_i's are unobservable, the e_i's are observable. Note that there will be n errors for each line, one error corresponding to every observation.

Similarly, there will be another set of n errors for another guessed line drawn through the data. For each guessed line, we can summarize its corresponding errors by one number, the sum of squares of these errors, which seems to be a natural criterion for penalizing a wrong guess. Note that a simple sum of these errors is not a good choice for a measure of misfit since positive errors end up canceling negative errors when both should be counted in our measure. However, this does not mean that the sum of squared error is the only single measure of misfit. Other measures include the sum of absolute errors, but this latter measure is mathematically more difficult to handle. Once the measure of misfit is chosen, α and β could then be estimated by minimizing this measure. In fact, this is the idea behind least squares estimation.

3.2 Least Squares Estimation and the Classical Assumptions

Least squares minimizes the residual sum of squares where the residuals are given by

$$e_i = Y_i - \widehat{\alpha} - \widehat{\beta}X_i \quad i = 1, 2, \ldots, n$$

and $\widehat{\alpha}$ and $\widehat{\beta}$ denote guesses on the regression parameters α and β, respectively. The residual sum of squares denoted by $RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - \widehat{\alpha} - \widehat{\beta}X_i)^2$ is minimized by the two first-order conditions:

$$\partial(\textstyle\sum_{i=1}^{n} e_i^2)/\partial\alpha = -2\sum_{i=1}^{n} e_i = 0; \text{ or } \sum_{i=1}^{n} Y_i - n\widehat{\alpha} - \widehat{\beta}\sum_{i=1}^{n} X_i = 0 \tag{3.2}$$

$$\partial(\textstyle\sum_{i=1}^{n} e_i^2)/\partial\beta = -2\sum_{i=1}^{n} e_i X_i = 0; \text{ or } \sum_{i=1}^{n} Y_i X_i - \widehat{\alpha}\sum_{i=1}^{n} X_i - \widehat{\beta}\sum_{i=1}^{n} X_i^2 = 0 \tag{3.3}$$

Solving the least squares normal equations given in (3.2) and (3.3) for α and β one gets

$$\widehat{\alpha}_{OLS} = \bar{Y} - \widehat{\beta}_{OLS}\bar{X} \text{ and } \widehat{\beta}_{OLS} = \textstyle\sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2 \tag{3.4}$$

where $\bar{Y} = \sum_{i=1}^{n} Y_i/n, \bar{X} = \sum_{i=1}^{n} X_i/n$, $y_i = Y_i - \bar{Y}$, $x_i = X_i - \bar{X}$, $\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$, $\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2$ and $\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$.

These estimators are subscripted by OLS denoting the ordinary least squares estimators. The OLS residuals $e_i = Y_i - \widehat{\alpha}_{OLS} - \widehat{\beta}_{OLS}X_i$ automatically satisfy the two numerical relationships given by (3.2) and (3.3). The first relationship states that (i) $\sum_{i=1}^{n} e_i = 0$, the residuals sum to zero. This is true as long as there is a constant in the regression. This numerical property of the least squares residuals also implies that the estimated regression line passes through the sample means $(\bar{X}, \bar{Y})$. To see this, average the residuals, or equation (3.2), this gives immediately $\bar{Y} = \widehat{\alpha}_{OLS} + \widehat{\beta}_{OLS}\bar{X}$. The second relationship states that (ii) $\sum_{i=1}^{n} e_i X_i = 0$, the residuals and the explanatory variable are uncorrelated. Other *numerical properties* that the OLS estimators satisfy are the following: (iii) $\sum_{i=1}^{n} \widehat{Y}_i = \sum_{i=1}^{n} Y_i$ and (iv) $\sum_{i=1}^{n} e_i \widehat{Y}_i = 0$. Property (iii) states that the sum of the estimated Y_i's or the predicted Y_i's from the sample is equal to the sum of the

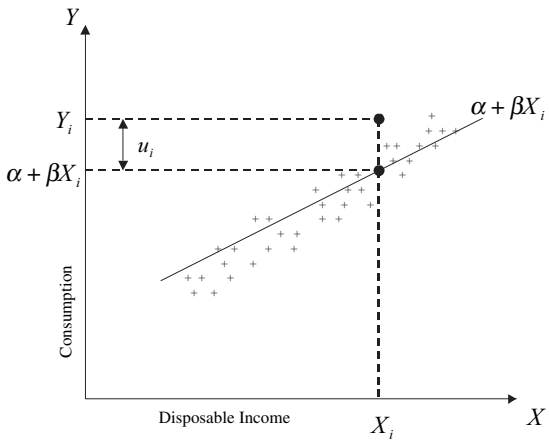


Figure 3.1 'True' Consumption Function

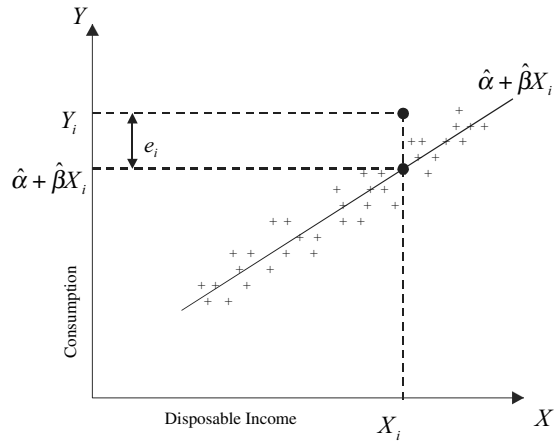


Figure 3.2 Estimated Consumption Function

actual Y_i's. Property (iv) states that the OLS residuals and the predicted Y_i's are uncorrelated. The proof of (iii) and (iv) *follow from (i) and (ii)* see problem 1. Of course, underlying our estimation of (3.1) is the assumption that (3.1) is the *true* model generating the data. In this case, (3.1) is *linear* in the parameters α and β, and contains only *one* explanatory variable X_i besides the constant. The inclusion of other explanatory variables in the model will be considered in Chapter 4, and the relaxation of the linearity assumption will be considered in Chapters 8 and 13. In order to study the statistical properties of the OLS estimators of α and β, we need to impose some statistical assumptions on the model generating the data.

Assumption 1: The disturbances have zero mean, i.e., $E(u_i) = 0$ for every $i = 1, 2, \ldots, n$. This assumption is needed to insure that on the average we are on the true line.

To see what happens if $E(u_i) \neq 0$, consider the case where households consistently under-report their consumption by a constant amount of δ dollars, while their income is measured accurately, say by cross-referencing it with their IRS tax forms. In this case,

$$(Observed\ Consumption) = (True\ Consumption) - \delta$$

and our regression equation is really

$$(True\ Consumption)_i = \alpha + \beta(Income)_i + u_i$$

But we observe,

$$(Observed\ Consumption)_i = \alpha + \beta(Income)_i + u_i - \delta$$

This can be thought of as the old regression equation with a new disturbance term $u_i^* = u_i - \delta$. Using the fact that $\delta > 0$ and $E(u_i) = 0$, one gets $E(u_i^*) = -\delta < 0$. This says that for all households with the same income, say $\$20,000$, their observed consumption will be on the average below that predicted from the true line $[\alpha + \beta(\$20,000)]$ by an amount δ. Fortunately, one

can deal with this problem of constant but non-zero mean of the disturbances by reparametizing the model as

$$(Observed\ Consumption)_i = \alpha^* + \beta(Income)_i + u_i$$

where $\alpha^* = \alpha - \delta$. In this case, $E(u_i) = 0$ and α^* and β can be estimated from the regression. Note that while α^* is estimable, α and δ are non-estimable. Also note that for all $\$20,000$ income households, their average consumption is $[(\alpha - \delta) + \beta(\$20,000)]$.

Assumption 2: The disturbances have a constant variance, i.e., $\text{var}(u_i) = \sigma^2$ for every $i = 1, 2, \ldots, n$. This insures that every observation is equally reliable.

To see what this assumption means, consider the case where $\text{var}(u_i) = \sigma_i^2$, for $i = 1, 2, \ldots, n$. In this case, each observation has a different variance. An observation with a large variance is less reliable than one with a smaller variance. But, how can this differing variance happen? In the case of consumption, households with large disposable income (a large X_i, say $\$100,000$) may be able to save more (or borrow more to spend more) than households with smaller income (a small X_i, say $\$10,000$). In this case, the variation in consumption for the $\$100,000$ income household will be much larger than that for the $\$10,000$ income household. Therefore, the corresponding variance for the $\$100,000$ observation will be larger than that for the $\$10,000$ observation. Consequences of different variances for different observations will be studied more rigorously in Chapter 5.

Assumption 3: The disturbances are not correlated, i.e., $E(u_i u_j) = 0$ for $i \neq j, i, j = 1, 2, \ldots, n$. Knowing the i-th disturbance does not tell us anything about the j-th disturbance, for $i \neq j$.

For the consumption example, the unforseen disturbance which caused the i-th household to consume more, (like a visit of a relative), has nothing to do with the unforseen disturbances of any other household. This is likely to hold for a random sample of households. However, it is less likely to hold for a time series study of consumption for the aggregate economy, where a disturbance in 1945, a war year, is likely to affect consumption for several years after that. In this case, we say that the disturbance in 1945 is related to the disturbances in 1946, 1947, and so on. Consequences of correlated disturbances will be studied in Chapter 5.

Assumption 4: The explanatory variable X is nonstochastic, i.e., fixed in repeated samples, and hence, not correlated with the disturbances. Also, $\sum_{i=1}^{n} x_i^2/n \neq 0$ and has a finite limit as n tends to infinity.

This assumption states that we have at least two distinct values for X. This makes sense, since we need at least two distinct points to draw a straight line. Otherwise $\bar{X} = X$, the common value, and $x = X - \bar{X} = 0$, which violates $\sum_{i=1}^{n} x_i^2 \neq 0$. In practice, one always has several distinct values of X. More importantly, this assumption implies that X is not a random variable and hence is not correlated with the disturbances.

In section 5.3, we will relax the assumption of a non-stochastic X. Basically, X becomes a random variable and our assumptions have to be recast *conditional* on the set of X's that are observed. This is the more realistic case with economic data. The zero mean assumption becomes $E(u_i/X) = 0$, the constant variance assumption becomes $\text{var}(u_i/X) = \sigma^2$, the no serial correlation assumption becomes $E(u_i u_j/X) = 0$ for $i \neq j$. The conditional expectation here is with respect to *every* observation on X_i from $i = 1, 2, \ldots n$. Of course, one can show that if $E(u_i/X) = 0$ for all i, then X_i and u_i are not correlated. The reverse is not necessarily true, see

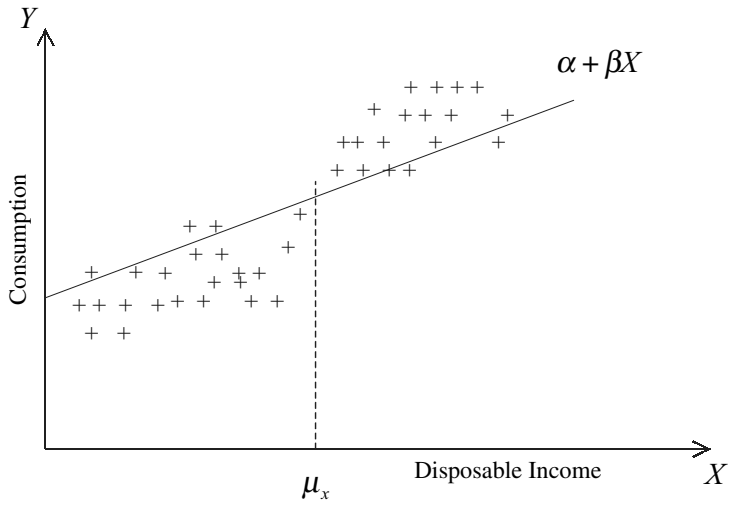


Figure 3.3 Consumption Function with $\text{Cov}(X, u) > 0$

problem 3 of Chapter 2. That problem shows that two random variables, say u_i and X_i could be uncorrelated, i.e., not linearly related when in fact they are nonlinearly related with $u_i = X_i^2$. Hence, $E(u_i/X_i) = 0$ is a stronger assumption than u_i and X_i are not correlated. By the law of iterated expectations given in the Appendix of Chapter 2, $E(u_i/X) = 0$ implies that $E(u_i) = 0$. It also implies that u_i is uncorrelated with *any* function of X_i. This is a stronger assumption than u_i is uncorrelated with X_i. Therefore, conditional on X_i, the mean of the disturbances is zero and does not depend on X_i. In this case, $E(Y_i/X_i) = \alpha + \beta X_i$ is linear in α and β and is assumed to be the *true* conditional mean of Y given X.

To see what a violation of assumption 4 means, suppose that X is a random variable and that X and u are positively correlated, then in the consumption example, households with income above the average income will be associated with disturbances above their mean of zero, and hence positive disturbances. Similarly, households with income below the average income will be associated with disturbances below their mean of zero, and hence negative disturbances. This means that the disturbances are systematically affected by values of the explanatory variable and the scatter of the data will look like Figure 3.3. Note that if we now erase the true line $(\alpha + \beta X)$, and estimate this line from the data, the least squares line drawn through the data is going to have a smaller intercept and a larger slope than those of the true line. The scatter should look like Figure 3.4 where the disturbances are random variables, not correlated with the X_i's, drawn from a distribution with zero mean and constant variance. Assumptions 1 and 4 insure that $E(Y_i/X_i) = \alpha + \beta X_i$, i.e., on the average we are on the true line. Several economic models will be studied where X and u are correlated. The consequences of this correlation will be studied in Chapters 5 and 11.

We now generate a data set which satisfies all four classical assumptions. Let α and β take the arbitrary values, say 10 and 0.5 respectively, and consider a set of 20 fixed X's, say income classes from \$10 to \$105 (in thousands of dollars), in increments of \$5, i.e., \$10, \$15, \$20, \$25, . . . , \$105. Our consumption variable Y_i is constructed as $(10 + 0.5X_i + u_i)$ where u_i is a

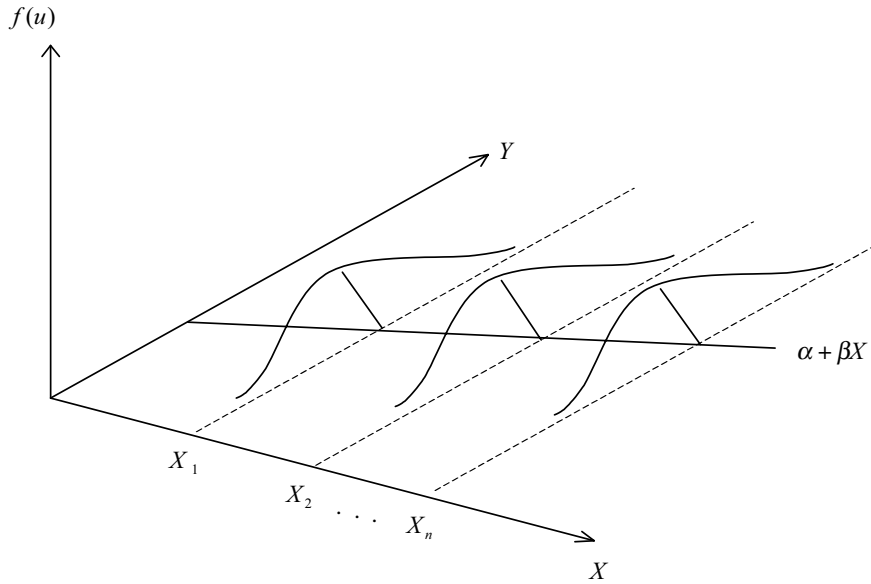


Figure 3.4 Random Disturbances around the Regression

disturbance which is a random draw from a distribution with zero mean and constant variance, say $\sigma^2 = 9$. Computers generate random numbers with various distributions.

In this case, Figure 3.4 would depict our data, with the true line being $(10 + 0.5X)$ and u_i being random draws from the computer which are by construction independent and identically distributed with mean zero and variance 9. For every set of 20 u_i's randomly generated, given the fixed X_i's, we obtain a corresponding set of 20 Y_i's from our linear regression model. This is what we mean in assumption 4 when we say that the X's are fixed in repeated samples. Monte Carlo experiments generate a large number of samples, say a 1000, in the fashion described above. For each data set generated, least squares can be performed and the properties of the resulting estimators which are derived analytically in the remainder of this chapter, can be verified. For example, the average of the 1000 estimates of α and β can be compared to their true values to see whether these least squares estimates are unbiased. Note what will happen to Figure 3.4 if $E(u_i) = -\delta$ where $\delta > 0$, or $\text{var}(u_i) = \sigma_i^2$ for $i = 1, 2, \ldots, n$. In the first case, the mean of $f(u)$, the probability density function of u, will shift off the true line $(10+0.5X)$ by $-\delta$. In other words, we can think of the distributions of the u_i's, shown in Figure 3.4 , being centered on a new imaginary line parallel to the true line but lower by a distance δ. This means that one is more likely to draw negative disturbances than positive disturbances, and the observed Y_i's are more likely to be below the true line than above it. In the second case, each $f(u_i)$ will have a different variance, hence the spread of this probability density function will vary with each observation. In this case, Figure 3.4 will have a distribution for the u_i's which has a different spread for each observation. In other words, if the u_i's are say normally distributed, then u_1 is drawn from a $N(0, \sigma_1^2)$ distribution, whereas u_2 is drawn from a $N(0, \sigma_2^2)$ distribution, and so on. Violation of the classical assumptions can also be studied using Monte Carlo experiments, see Chapter 5.

3.3 Statistical Properties of Least Squares

(i) Unbiasedness

Given assumptions 1-4, it is easy to show that $\widehat{\beta}_{OLS}$ is unbiased for β. In fact, using equation (3.4) one can write

$$\widehat{\beta}_{OLS} = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i Y_i / \sum_{i=1}^{n} x_i^2 = \beta + \sum_{i=1}^{n} x_i u_i / \sum_{i=1}^{n} x_i^2 \tag{3.5}$$

where the second equality follows from the fact that $y_i = Y_i - \bar{Y}$ and $\sum_{i=1}^{n} x_i \bar{Y} = \bar{Y} \sum_{i=1}^{n} x_i = 0$. The third equality follows from substituting Y_i from (3.1) and using the fact that $\sum_{i=1}^{n} x_i = 0$. Taking expectations of both sides of (3.5) and using assumptions 1 and 4, one can show that $E(\widehat{\beta}_{OLS}) = \beta$. Furthermore, one can derive the variance of $\widehat{\beta}_{OLS}$ from (3.5) since

$$\begin{aligned} \text{var}(\widehat{\beta}_{OLS}) &= E(\widehat{\beta}_{OLS} - \beta)^2 = E(\textstyle\sum_{i=1}^{n} x_i u_i / \sum_{i=1}^{n} x_i^2)^2 \\ &= \text{var}(\textstyle\sum_{i=1}^{n} x_i u_i / \sum_{i=1}^{n} x_i^2) = \sigma^2 / \sum_{i=1}^{n} x_i^2 \end{aligned} \tag{3.6}$$

where the last equality uses assumptions 2 and 3, i.e., that the u_i's are not correlated with each other and that their variance is constant, see problem 4. Note that the variance of the OLS estimator of β depends upon σ^2, the variance of the disturbances in the true model, and on the variation in X. The larger the variation in X the larger is $\sum_{i=1}^{n} x_i^2$ and the smaller is the variance of $\widehat{\beta}_{OLS}$.

(ii) Consistency

Next, we show that $\widehat{\beta}_{OLS}$ is consistent for β. A sufficient condition for consistency is that $\widehat{\beta}_{OLS}$ is unbiased and its variance tends to zero as n tends to infinity. We have already shown $\widehat{\beta}_{OLS}$ to be unbiased, it remains to show that its variance tends to zero as n tends to infinity.

$$\lim_{n\to\infty} \text{var}(\widehat{\beta}_{OLS}) = \lim_{n\to\infty} [(\sigma^2/n)/(\textstyle\sum_{i=1}^{n} x_i^2/n)] = 0$$

where the second equality follows from the fact that $(\sigma^2/n) \to 0$ and $(\sum_{i=1}^{n} x_i^2/n) \neq 0$ and has a finite limit, see assumption 4. Hence, plim $\widehat{\beta}_{OLS} = \beta$ and $\widehat{\beta}_{OLS}$ is consistent for β. Similarly one can show that $\widehat{\alpha}_{OLS}$ is unbiased and consistent for α with variance $\sigma^2 \sum_{i=1}^{n} X_i^2 / n \sum_{i=1}^{n} x_i^2$, and $\text{cov}(\widehat{\alpha}_{OLS}, \widehat{\beta}_{OLS}) = -\bar{X}\sigma^2 / \sum_{i=1}^{n} x_i^2$, see problem 5.

(iii) Best Linear Unbiased

Using (3.5) one can write $\widehat{\beta}_{OLS}$ as $\sum_{i=1}^{n} w_i Y_i$ where $w_i = x_i / \sum_{i=1}^{n} x_i^2$. This proves that $\widehat{\beta}_{OLS}$ is a linear combination of the Y_i's, with weights w_i satisfying the following properties:

$$\textstyle\sum_{i=1}^{n} w_i = 0; \sum_{i=1}^{n} w_i X_i = 1; \sum_{i=1}^{n} w_i^2 = 1 / \sum_{i=1}^{n} x_i^2 \tag{3.7}$$

The next theorem shows that among all linear unbiased estimators of β, it is $\widehat{\beta}_{OLS}$ which has the smallest variance. This is known as the Gauss-Markov Theorem.

Theorem 1: Consider any arbitrary linear estimator $\widetilde{\beta} = \sum_{i=1}^{n} a_i Y_i$ for β, where the a_i's denote arbitrary constants. If $\widetilde{\beta}$ is unbiased for β, and assumptions 1 to 4 are satisfied, then $\text{var}(\widetilde{\beta}) \geq \text{var}(\widehat{\beta}_{OLS})$.

Proof: Substituting Y_i from (3.1) into $\widetilde{\beta}$, one gets $\widetilde{\beta} = \alpha \sum_{i=1}^n a_i + \beta \sum_{i=1}^n a_i X_i + \sum_{i=1}^n a_i u_i$. For $\widetilde{\beta}$ to be unbiased for β it must follow that $E(\widetilde{\beta}) = \alpha \sum_{i=1}^n a_i + \beta \sum_{i=1}^n a_i X_i = \beta$ for all observations $i = 1, 2, \ldots, n$. This means that $\sum_{i=1}^n a_i = 0$ and $\sum_{i=1}^n a_i X_i = 1$ for all $i = 1, 2, \ldots, n$. Hence, $\widetilde{\beta} = \beta + \sum_{i=1}^n a_i u_i$ with $\text{var}(\widetilde{\beta}) = \text{var}(\sum_{i=1}^n a_i u_i) = \sigma^2 \sum_{i=1}^n a_i^2$ where the last equality follows from assumptions 2 and 3. But the a_i's are constants which differ from the w_i's, the weights of the OLS estimator, by some other constants, say d_i's, i.e., $a_i = w_i + d_i$ for $i = 1, 2, \ldots, n$. Using the properties of the a_i's and w_i one can deduce similar properties on the d_i's i.e., $\sum_{i=1}^n d_i = 0$ and $\sum_{i=1}^n d_i X_i = 0$. In fact,

$$\sum_{i=1}^n a_i^2 = \sum_{i=1}^n d_i^2 + \sum_{i=1}^n w_i^2 + 2 \sum_{i=1}^n w_i d_i$$

where $\sum_{i=1}^n w_i d_i = \sum_{i=1}^n x_i d_i / \sum_{i=1}^n x_i^2 = 0$. This follows from the definition of w_i and the fact that $\sum_{i=1}^n d_i = \sum_{i=1}^n d_i X_i = 0$. Hence,

$$\text{var}(\widetilde{\beta}) = \sigma^2 \sum_{i=1}^n a_i^2 = \sigma^2 \sum_{i=1}^n d_i^2 + \sigma^2 \sum_{i=1}^n w_i^2 = \text{var}(\widehat{\beta}_{OLS}) + \sigma^2 \sum_{i=1}^n d_i^2$$

Since $\sigma^2 \sum_{i=1}^n d_i^2$ is non-negative, this proves that $\text{var}(\widetilde{\beta}) \geq \text{var}(\widehat{\beta}_{OLS})$ with the equality holding only if $d_i = 0$ for all $i = 1, 2, \ldots, n$, i.e., only if $a_i = w_i$, in which case $\widetilde{\beta}$ reduces to $\widehat{\beta}_{OLS}$. Therefore, any linear estimator of β, like $\widetilde{\beta}$ that is unbiased for β has variance at least as large as $\text{var}(\widehat{\beta}_{OLS})$. This proves that $\widehat{\beta}_{OLS}$ is BLUE, Best among all Linear Unbiased Estimators of β.

Similarly, one can show that $\widehat{\alpha}_{OLS}$ is linear in Y_i and has the smallest variance among all linear unbiased estimators of α, if assumptions 1 to 4 are satisfied, see problem 6. This result implies that the OLS estimator of α is also BLUE.

3.4 Estimation of σ^2

The variance of the regression disturbances σ^2 is unknown and has to be estimated. In fact, both the variance of $\widehat{\beta}_{OLS}$ and that of $\widehat{\alpha}_{OLS}$ depend upon σ^2, see (3.6) and problem 5. An unbiased estimator for σ^2 is $s^2 = \sum_{i=1}^n e_i^2/(n-2)$. To prove this, we need the fact that

$$e_i = Y_i - \widehat{\alpha}_{OLS} - \widehat{\beta}_{OLS} X_i = y_i - \widehat{\beta}_{OLS} x_i = (\beta - \widehat{\beta}_{OLS}) x_i + (u_i - \bar{u})$$

where $\bar{u} = \sum_{i=1}^n u_i/n$. The second equality substitutes $\widehat{\alpha}_{OLS} = \bar{Y} - \widehat{\beta}_{OLS} \bar{X}$ and the third equality substitutes $y_i = \beta x_i + (u_i - \bar{u})$. Hence,

$$\sum_{i=1}^n e_i^2 = (\widehat{\beta}_{OLS} - \beta)^2 \sum_{i=1}^n x_i^2 + \sum_{i=1}^n (u_i - \bar{u})^2 - 2(\widehat{\beta}_{OLS} - \beta) \sum_{i=1}^n x_i (u_i - \bar{u}),$$

and

$$\begin{aligned} E(\textstyle\sum_{i=1}^n e_i^2) &= \textstyle\sum_{i=1}^n x_i^2 \text{var}(\widehat{\beta}_{OLS}) + (n-1)\sigma^2 - 2E(\sum_{i=1}^n x_i u_i)^2 / \sum_{i=1}^n x_i^2 \\ &= \sigma^2 + (n-1)\sigma^2 - 2\sigma^2 = (n-2)\sigma^2 \end{aligned}$$

where the first equality uses the fact that $E(\sum_{i=1}^n (u_i - \bar{u})^2) = (n-1)\sigma^2$ and $\widehat{\beta}_{OLS} - \beta = \sum_{i=1}^n x_i u_i / \sum_{i=1}^n x_i^2$. The second equality uses the fact that $\text{var}(\widehat{\beta}_{OLS}) = \sigma^2 / \sum_{i=1}^n x_i^2$ and

$$E(\textstyle\sum_{i=1}^n x_i u_i)^2 = \sigma^2 \sum_{i=1}^n x_i^2.$$

Therefore, $E(s^2) = E(\sum_{i=1}^n e_i^2/(n-2)) = \sigma^2$.

Intuitively, the estimator of σ^2 could be obtained from $\sum_{i=1}^n (u_i - \bar{u})^2/(n-1)$ if the true disturbances were known. Since the u's are not known, consistent estimates of them are used. These are the e_i's. Since $\sum_{i=1}^n e_i = 0$, our estimator of σ^2 becomes $\sum_{i=1}^n e_i^2/(n-1)$. Taking expectations we find that the correct divisor ought to be $(n-2)$ and not $(n-1)$ for this estimator to be unbiased for σ^2. This is plausible, since we have estimated two parameters α and β in obtaining the e_i's, and there are only $n-2$ independent pieces of information left in the data. To prove this fact, consider the OLS normal equations given in (3.2) and (3.3). These equations represent two relationships involving the e_i's. Therefore, knowing $(n-2)$ of the e_i's we can deduce the remaining two e_i's from (3.2) and (3.3).

3.5 Maximum Likelihood Estimation

Assumption 5: The u_i's are independent and identically distributed $N(0, \sigma^2)$.

This assumption allows us to derive distributions of estimators and other test statistics. In fact using (3.5) one can easily see that $\widehat{\beta}_{OLS}$ is a linear combination of the u_i's. But, a linear combination of normal random variables is itself a normal random variable, see Chapter 2, problem 15. Hence, $\widehat{\beta}_{OLS}$ is $N(\beta, \sigma^2/\sum_{i=1}^n x_i^2)$. Similarly $\widehat{\alpha}_{OLS}$ is $N(\alpha, \sigma^2 \sum_{i=1}^n X_i^2/n \sum_{i=1}^n x_i^2)$, and Y_i is $N(\alpha+\beta X_i, \sigma^2)$. Moreover, we can write the joint probability density function of the u's as $f(u_1, u_2, \ldots, u_n; \alpha, \beta, \sigma^2) = (1/2\pi\sigma^2)^{n/2} \exp(-\sum_{i=1}^n u_i^2/2\sigma^2)$. To get the likelihood function we make the transformation $u_i = Y_i - \alpha - \beta X_i$ and note that the Jacobian of the transformation is 1. Therefore,

$$f(Y_1, Y_2, \ldots, Y_n; \alpha, \beta, \sigma^2) = (1/2\pi\sigma^2)^{n/2} \exp\{-\textstyle\sum_{i=1}^n (Y_i - \alpha - \beta X_i)^2/2\sigma^2\} \tag{3.8}$$

Taking the log of this likelihood, we get

$$\log L(\alpha, \beta, \sigma^2) = -(n/2)\log(2\pi\sigma^2) - \textstyle\sum_{i=1}^n (Y_i - \alpha - \beta X_i)^2/2\sigma^2 \tag{3.9}$$

Maximizing this likelihood with respect to α, β and σ^2 one gets the maximum likelihood estimators (MLE). However, only the second term in the log likelihood contains α and β and that term (without the negative sign) has already been minimized with respect to α and β in (3.2) and (3.3) giving us the OLS estimators. Hence, $\widehat{\alpha}_{MLE} = \widehat{\alpha}_{OLS}$ and $\widehat{\beta}_{MLE} = \widehat{\beta}_{OLS}$. Similarly, by differentiating $\log L$ with respect to σ^2 and setting this derivative equal to zero one gets $\widehat{\sigma}^2_{MLE} = \sum_{i=1}^n e_i^2/n$, see problem 7. Note that this differs from s^2 only in the divisor. In fact, $E(\widehat{\sigma}^2_{MLE}) = (n-2)\sigma^2/n \neq \sigma^2$. Hence, $\widehat{\sigma}^2_{MLE}$ is biased but note that it is still asymptotically unbiased.

So far, the gains from imposing assumption 5 are the following: The likelihood can be formed, maximum likelihood estimators can be derived, and distributions can be obtained for these estimators. One can also derive the Cramér-Rao lower bound for unbiased estimators of the parameters and show that the $\widehat{\alpha}_{OLS}$ and $\widehat{\beta}_{OLS}$ attain this bound whereas s^2 does not. This derivation is postponed until Chapter 7. In fact, one can show following the theory of complete sufficient statistics that $\widehat{\alpha}_{OLS}$, $\widehat{\beta}_{OLS}$ and s^2 are *minimum variance unbiased* estimators for α, β and σ^2, see Chapter 2. This is a stronger result (for $\widehat{\alpha}_{OLS}$ and $\widehat{\beta}_{OLS}$) than that obtained using the Gauss-Markov Theorem. It says, that among all unbiased estimators of α and β, the OLS

estimators are the best. In other words, our set of estimators include now *all* unbiased estimators and not just *linear* unbiased estimators. This stronger result is obtained at the expense of a stronger distributional assumption, i.e., normality. If the distribution of the disturbances is not normal, then OLS is no longer MLE. In this case, MLE will be more efficient than OLS as long as the distribution of the disturbances is correctly specified. Some of the advantages and disadvantages of MLE were discussed in Chapter 2.

We found the distributions of $\widehat{\alpha}_{OLS}$, $\widehat{\beta}_{OLS}$, now we give that of s^2. In Chapter 7, it is shown that $\sum_{i=1}^{n} e_i^2/\sigma^2$ is a chi-squared with $(n-2)$ degrees of freedom. Also, s^2 is independent of $\widehat{\alpha}_{OLS}$ and $\widehat{\beta}_{OLS}$. This is useful for test of hypotheses. In fact, the major gain from assumption 5 is that we can perform test of hypotheses.

Standardizing the normal random variable $\widehat{\beta}_{OLS}$, one gets $z = (\widehat{\beta}_{OLS} - \beta)/(\sigma^2/\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}} \sim N(0,1)$. Also, $(n-2)s^2/\sigma^2$ is distributed as χ^2_{n-2}. Hence, one can divide z, a $N(0,1)$ random variable, by the square root of $(n-2)s^2/\sigma^2$ divided by its degrees of freedom $(n-2)$ to get a t-statistic with $(n-2)$ degrees of freedom. The resulting statistic is $t_{obs} = (\widehat{\beta}_{OLS} - \beta)/(s^2/\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}} \sim t_{n-2}$, see problem 8. This statistic can be used to test $H_0; \beta = \beta_0$, versus $H_1; \beta \neq \beta_0$, where β_0 is a known constant. Under H_0, t_{obs} can be calculated and its value can be compared to a critical value from a t-distribution with $(n-2)$ degrees of freedom, at a specified critical value of $\alpha\%$. Of specific interest is the hypothesis $H_0; \beta = 0$, which states that there is no linear relationship between Y_i and X_i. Under H_0,

$$t_{obs} = \widehat{\beta}_{OLS}/(s^2/\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}} = \widehat{\beta}_{OLS}/\widehat{se}(\widehat{\beta}_{OLS})$$

where $\widehat{se}(\widehat{\beta}_{OLS}) = (s^2/\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}}$. If $|t_{obs}| > t_{\alpha/2;n-2}$, then H_0 is rejected at the $\alpha\%$ significance level. $t_{\alpha/2;n-2}$ represents a critical value obtained from a t-distribution with $n-2$ degrees of freedom. It is determined such that the area to its right under a t_{n-2} distribution is equal to $\alpha/2$.

Similarly one can get a confidence interval for β by using the fact that, $\Pr[-t_{\alpha/2;n-2} < t_{obs} < t_{\alpha/2;n-2}] = 1-\alpha$ and substituting for t_{obs} its value derived above as $(\widehat{\beta}_{OLS} - \beta)/\widehat{se}(\widehat{\beta}_{OLS})$. Since the critical values are known, $\widehat{\beta}_{OLS}$ and $\widehat{se}(\widehat{\beta}_{OLS})$ can be calculated from the data, the following $(1-\alpha)\%$ confidence interval for β emerges

$$\widehat{\beta}_{OLS} \pm t_{\alpha/2;n-2}\widehat{se}(\widehat{\beta}_{OLS}).$$

Tests of hypotheses and confidence intervals on α and σ^2 can be similarly constructed using the normal distribution of $\widehat{\alpha}_{OLS}$ and the χ^2_{n-2} distribution of $(n-2)s^2/\sigma^2$.

3.6 A Measure of Fit

We have obtained the least squares estimates of α, β and σ^2 and found their distributions under normality of the disturbances. We have also learned how to test hypotheses regarding these parameters. Now we turn to a measure of fit for this estimated regression line. Recall, that $e_i = Y_i - \widehat{Y}_i$ where $\widehat{Y}_i$ denotes the predicted Y_i from the least squares regression line at the value X_i, i.e., $\widehat{\alpha}_{OLS} + \widehat{\beta}_{OLS}X_i$. Using the fact that $\sum_{i=1}^{n} e_i = 0$, we deduce that $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \widehat{Y}_i$, and therefore, $\bar{Y} = \overline{\widehat{Y}}$. The actual and predicted values of Y have the same sample mean, see numerical properties (i) and (iii) of the OLS estimators discussed in section 2. This is true

as long as there is a constant in the regression. Adding and subtracting $\bar{Y}$ from e_i, we get $e_i = y_i - \widehat{y}_i$, or $y_i = e_i + \widehat{y}_i$. Squaring and summing both sides:

$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^n e_i^2 + \sum_{i=1}^n \widehat{y}_i^2 + 2\sum_{i=1}^n e_i\widehat{y}_i = \sum_{i=1}^n e_i^2 + \sum_{i=1}^n \widehat{y}_i^2 \tag{3.10}$$

where the last equality follows from the fact that $\widehat{y}_i = \widehat{\beta}_{OLS}x_i$ and $\sum_{i=1}^n e_i x_i = 0$. In fact,

$$\sum_{i=1}^n e_i\widehat{y}_i = \sum_{i=1}^n e_i\widehat{Y}_i = 0$$

means that the OLS residuals are uncorrelated with the predicted values from the regression, see numerical properties (ii) and (iv) of the OLS estimates discussed in section 3.2. In other words, (3.10) says that the total variation in Y_i, around its sample mean $\bar{Y}$ i.e., $\sum_{i=1}^n y_i^2$, can be decomposed into two parts: the first is the regression sums of squares $\sum_{i=1}^n \widehat{y}_i^2 = \widehat{\beta}_{OLS}^2 \sum_{i=1}^n x_i^2$, and the second is the residual sums of squares $\sum_{i=1}^n e_i^2$. In fact, regressing Y on a constant yields $\widetilde{\alpha}_{OLS} = \bar{Y}$, see problem 2, and the unexplained residual sums of squares of this naive model is

$$\sum_{i=1}^n (Y_i - \widetilde{\alpha}_{OLS})^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n y_i^2.$$

Therefore, $\sum_{i=1}^n \widehat{y}_i^2$ in (3.10) gives the explanatory power of X after the constant is fit.

Using this decomposition, one can define the explanatory power of the regression as the ratio of the regression sums of squares to the total sums of squares. In other words, define $R^2 = \sum_{i=1}^n \widehat{y}_i^2 / \sum_{i=1}^n y_i^2$ and this value is clearly between 0 and 1. In fact, dividing (3.10) by $\sum_{i=1}^n y_i^2$ one gets $R^2 = 1 - \sum_{i=1}^n e_i^2 / \sum_{i=1}^n y_i^2$. The $\sum_{i=1}^n e_i^2$ is a measure of misfit which was minimized by least squares. If $\sum_{i=1}^n e_i^2$ is large, this means that the regression is not explaining a lot of the variation in Y and hence, the R^2 value would be small. Alternatively, if the $\sum_{i=1}^n e_i^2$ is small, then the fit is good and R^2 is large. In fact, for a perfect fit, where all the observations lie on the fitted line, $Y_i = \widehat{Y}_i$ and $e_i = 0$, which means that $\sum_{i=1}^n e_i^2 = 0$ and $R^2 = 1$. The other extreme case is where the regression sums of squares $\sum_{i=1}^n \widehat{y}_i^2 = 0$. In other words, the linear regression explains nothing of the variation in Y_i. In this case, $\sum_{i=1}^n y_i^2 = \sum_{i=1}^n e_i^2$ and $R^2 = 0$. Note that since $\sum_{i=1}^n \widehat{y}_i^2 = 0$ implies $\widehat{y}_i = 0$ for every i, which in turn means that $\widehat{Y}_i = \bar{Y}$ for every i. The fitted regression line is a horizontal line drawn at $Y = \bar{Y}$, and the independent variable X does not have any explanatory power in a linear relationship with Y.

Note that R^2 has two alternative meanings: (i) It is the simple squared correlation coefficient between Y_i and $\widehat{Y}_i$, see problem 9. Also, for the simple regression case, (ii) it is the simple squared correlation between X and Y. This means that before one runs the regression of Y on X, one can compute r_{xy}^2 which in turn tells us the proportion of the variation in Y that will be explained by X. If this number is pretty low, we have a weak linear relationship between Y and X and we know that a poor fit will result if Y is regressed on X. It is worth emphasizing that R^2 is a measure of *linear* association between Y and X. There could exist, for example, a perfect quadratic relationship between X and Y, yet the estimated least squares line through the data is a flat line implying that $R^2 = 0$, see problem 3 of Chapter 2. One should also be suspicious of least squares regressions with R^2 that are too close to 1. In some cases, we may not want to include a constant in the regression. In such cases, one should use an *uncentered* R^2 as a measure fit. The appendix to this chapter defines both *centered* and *uncentered* R^2 and explains the difference between them.

3.7 Prediction

Let us now predict Y_0 given X_0. Usually this is done for a time series regression, where the researcher is interested in predicting the future, say one period ahead. This new observation Y_0 is generated by (3.1), i.e.,

$$Y_0 = \alpha + \beta X_0 + u_0 \tag{3.11}$$

What is the Best Linear Unbiased Predictor (BLUP) of $E(Y_0)$? From (3.11), $E(Y_0) = \alpha + \beta X_0$ is a linear combination of α and β. Using the Gauss-Markov result, $\widehat{Y}_0 = \widehat{\alpha}_{OLS} + \widehat{\beta}_{OLS} X_0$ is BLUE for $\alpha + \beta X_0$ and the variance of this predictor of $E(Y_0)$ is $\sigma^2[(1/n) + (X_0 - \bar{X})^2 / \sum_{i=1}^n x_i^2]$, see problem 10. But, what if we are interested in the BLUP for Y_0 itself ? Y_0 differs from $E(Y_0)$ by u_0, and the best predictor of u_0 is zero, so the BLUP for Y_0 is still $\widehat{Y}_0$. The forecast error is

$$Y_0 - \widehat{Y}_0 = [Y_0 - E(Y_0)] + [E(Y_0) - \widehat{Y}_0] = u_0 + [E(Y_0) - \widehat{Y}_0]$$

where u_0 is the error committed even if the true regression line is known, and $E(Y_0) - \widehat{Y}_0$ is the difference between the sample and population regression lines. Hence, the variance of the forecast error becomes:

$$\text{var}(u_0) + \text{var}[E(Y_0) - \widehat{Y}_0] + 2\text{cov}[u_0, E(Y_0) - \widehat{Y}_0] = \sigma^2[1 + (1/n) + (X_0 - \bar{X})^2 / \sum_{i=1}^n x_i^2]$$

This says that the variance of the forecast error is equal to the variance of the predictor of $E(Y_0)$ plus the var(u_0) plus twice the covariance of the predictor of $E(Y_0)$ and u_0. But, this last covariance is zero, since u_0 is a new disturbance and is not correlated with the disturbances in the sample upon which $\widehat{Y}_i$ is based. Therefore, the predictor of the average consumption of a $\$20,000$ income household is the same as the predictor of consumption for a specific household whose income is $\$20,000$. The difference is not in the predictor itself but in the variance attached to it. The latter variance being larger only by σ^2, the variance of u_0. The variance of the predictor therefore, depends upon σ^2, the sample size, the variation in the X's, and how far X_0 is from the sample mean of the observed data. To summarize, the smaller σ^2 is, the larger n and $\sum_{i=1}^n x_i^2$ are, and the closer X_0 is to $\bar{X}$, the smaller is the variance of the predictor. One can construct 95% confidence intervals to these predictions for every value of X_0. In fact, this is $(\widehat{\alpha}_{OLS} + \widehat{\beta}_{OLS} X_0) \pm t_{.025;n-2}\{s[1 + (1/n) + (X_0 - \bar{X})^2 / \sum_{i=1}^n x_i^2]^{\frac{1}{2}}\}$ where s replaces σ, and $t_{.025;n-2}$ represents the 2.5% critical value obtained from a t-distribution with $n-2$ degrees of freedom. Figure 3.5 shows this confidence band around the estimated regression line. This is a hyperbola which is the narrowest at $\bar{X}$ as expected, and widens as we predict away from $\bar{X}$.

3.8 Residual Analysis

A plot of the residuals of the regression is very important. The residuals are consistent estimates of the true disturbances. But unlike the u_i's, these e_i's are not independent. In fact, the OLS normal equations (3.2)and (3.3) give us two relationships between these residuals. Therefore, knowing $(n-2)$ of these residuals the remaining two residuals can be deduced. If we had the true u_i's, and we plotted them, they should look like a random scatter around the horizontal axis with no specific pattern to them. A plot of the e_i's that shows a certain pattern like a set of positive residuals followed by a set of negative residuals as shown in Figure 3.6(a) may be

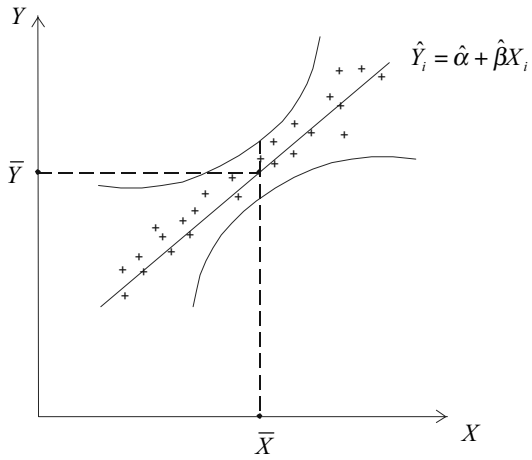


Figure 3.5 95% Confidence Bands

indicative of a violation of one of the 5 assumptions imposed on the model, or simply indicating a wrong functional form. For example, if assumption 3 is violated, so that the u_i's are say positively correlated, then it is likely to have a positive residual followed by a positive one, and a negative residual followed by a negative one, as observed in Figure 3.6(b). Alternatively, if we fit a linear regression line to a true quadratic relation between Y and X, then a scatter of residuals like that in Figure 3.6(c) will be generated. We will study how to deal with this violation and how to test for it in Chapter 5.

Large residuals are indicative of bad predictions in the sample. A large residual could be a typo, where the researcher entered this observation wrongly. Alternatively, it could be an influential observation, or an outlier which behaves differently from the other data points in the sample and therefore, is further away from the estimated regression line than the other data points. The fact that OLS minimizes the sum of squares of these residuals means that a large weight is put on this observation and hence it is influential. In other words, removing this observation from the sample may change the estimates and the regression line significantly. For more on the study of influential observations, see Belsely, Kuh and Welsch (1980). We will focus on this issue in Chapter 8 of this book.

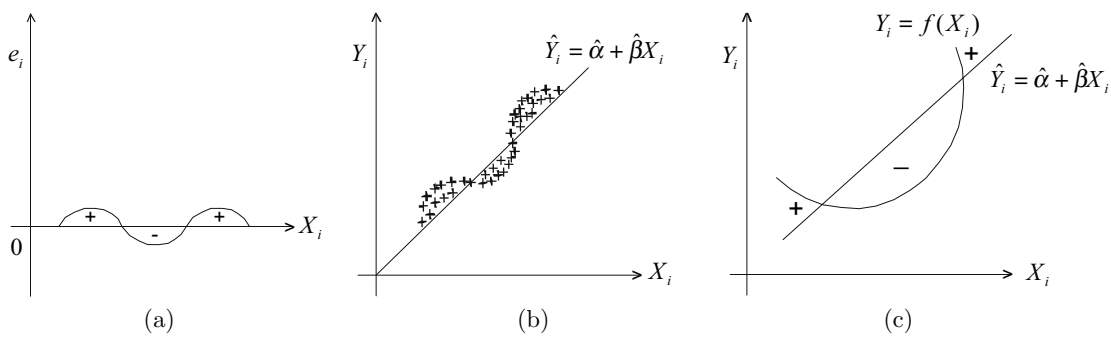


Figure 3.6 Positively Correlated Residuals

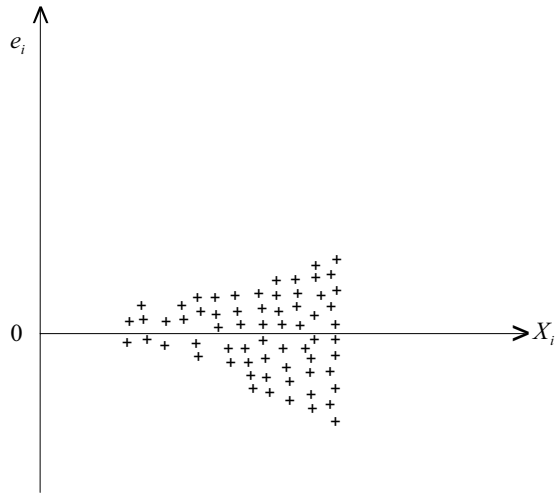


Figure 3.7 Residual Variation Growing with X

One can also plot the residuals versus the X_i's. If a pattern like Figure 3.7 emerges, this could be indicative of a violation of assumption 2 because the variation of the residuals is growing with X_i when it should be constant for all observations. Alternatively, it could imply a relationship between the X_i's and the true disturbances which is a violation of assumption 4.

In summary, one should always plot the residuals to check the data, identify influential observations, and check violations of the 5 assumptions underlying the regression model. In the next few chapters, we will study various tests of the violation of the classical assumptions. Most of these tests are based on the residuals of the model. These tests along with residual plots should help the researcher gauge the adequacy of his or her model.

Table 3.1 Simple Regression Computations

OBS	Consumption y_i	Income x_i	y_i=Y_i–$\bar{Y}$	x_i=X_i-$\bar{X}$	$x_i y_i$	x_i^2	$\hat{Y}_i$	e_i
1	4.6	5	–1.9	–2.5	4.75	6.25	4.476190	0.123810
2	3.6	4	–2.9	–3.5	10.15	12.25	3.666667	–0.066667
3	4.6	6	–1.9	–1.5	2.85	2.25	5.285714	–0.685714
4	6.6	8	0.1	0.5	0.05	0.25	6.904762	–0.304762
5	7.6	8	1.1	0.5	0.55	0.25	6.904762	0.695238
6	5.6	7	–0.9	–0.5	0.45	0.25	6.095238	–0.495238
7	5.6	6	–0.9	–1.5	1.35	2.25	5.285714	0.314286
8	8.6	9	2.1	1.5	3.15	2.25	7.714286	0.885714
9	8.6	10	2.1	2.5	5.25	6.25	8.523810	0.076190
10	9.6	12	3.1	4.5	13.95	20.25	10.142857	–0.542857
SUM	6.5	75	0	0	42.5	52.5	65	0
MEAN	6.5	7.5					6.5	

3.9 Numerical Example

Table 3.1 gives the annual consumption of 10 households each selected randomly from a group of households with a fixed personal disposable income. Both income and consumption are measured in \$10,000, so that the first household earns \$50,000 and consumes \$46,000 annually. It is worthwhile doing the computations necessary to obtain the least squares regression estimates of consumption on income in this simple case and to compare them with those obtained from a regression package. In order to do this, we first compute $\bar{Y} = 6.5$ and $\bar{X} = 7.5$ and form two new columns of data made up of $y_i = Y_i - \bar{Y}$ and $x_i = X_i - \bar{X}$. To get $\widehat{\beta}_{OLS}$ we need $\sum_{i=1}^{n} x_i y_i$, so we multiply these last two columns by each other and sum to get 42.5. The denominator of $\widehat{\beta}_{OLS}$ is given by $\sum_{i=1}^{n} x_i^2$. This is why we square the x_i column to get x_i^2 and sum to obtain 52.5. Our estimate of $\widehat{\beta}_{OLS} = 42.5/52.5 = 0.8095$ which is the estimated marginal propensity to consume. This is the extra consumption brought about by an extra dollar of disposable income.

$$\widehat{\alpha}_{OLS} = \bar{Y} - \widehat{\beta}_{OLS}\bar{X} = 6.5 - (0.8095)(7.5) = 0.4286$$

This is the estimated consumption at zero personal disposable income. The fitted values or predicted values from this regression are computed from $\widehat{Y}_i = \widehat{\alpha}_{OLS} + \widehat{\beta}_{OLS}X_i = 0.4286 + 0.8095X_i$ and are given in Table 3.1. Note that the mean of $\widehat{Y}_i$ is equal to the mean of Y_i confirming one of the numerical properties of least squares. The residuals are computed from $e_i = Y_i - \widehat{Y}_i$ and they satisfy $\sum_{i=1}^{n} e_i = 0$. It is left to the reader to verify that $\sum_{i=1}^{n} e_i X_i = 0$. The residual sum of squares is obtained by squaring the column of residuals and summing it. This gives us $\sum_{i=1}^{n} e_i^2 = 2.495238$. This means that $s^2 = \sum_{i=1}^{n} e_i^2/(n-2) = 0.311905$. Its square root is given by $s = 0.558$. This is known as the standard error of the regression. In this case, the estimated $\text{var}(\widehat{\beta}_{OLS})$ is $s^2/\sum_{i=1}^{n} x_i^2 = 0.311905/52.5 = 0.005941$ and the estimated

$$\text{var}(\widehat{\alpha}) = s^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} x_i^2}\right] = 0.311905\left[\frac{1}{10} + \frac{(7.5)^2}{52.5}\right] = 0.365374$$

Taking the square root of these estimated variances, we get the estimated standard errors of $\widehat{\alpha}_{OLS}$ and $\widehat{\beta}_{OLS}$ given by $\widehat{se}(\widehat{\alpha}_{OLS}) = 0.60446$ and $\widehat{se}(\widehat{\beta}_{OLS}) = 0.077078$.

Since the disturbances are normal, the OLS estimators are also the maximum likelihood estimators, and are normally distributed themselves. For the null hypothesis $H_0^a; \beta = 0$; the observed t-statistic is

$$t_{obs} = (\widehat{\beta}_{OLS} - 0)/\widehat{se}(\widehat{\beta}_{OLS}) = 0.809524/0.077078 = 10.50$$

and this is highly significant, since $\Pr[|t_8| > 10.5] < 0.0001$. This probability can be obtained using most regression packages. It is also known as the p-value or probability value. It shows that this t-value is highly unlikely and we reject H_0^a that $\beta = 0$. Similarly, the null hypothesis $H_0^b; \alpha = 0$, gives an observed t-statistic of $t_{obs} = (\widehat{\alpha}_{OLS} - 0)/\widehat{se}(\widehat{\alpha}_{OLS}) = 0.428571/0.604462 = 0.709$, which is not significant, since its p-value is $\Pr[|t_8| > 0.709] < 0.498$. Hence, we do not reject the null hypothesis H_0^b that $\alpha = 0$.

The total sum of squares is $\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ which can be obtained by squaring the y_i column in Table 3.1 and summing. This yields $\sum_{i=1}^{n} y_i^2 = 36.9$. Also, the regression sum of squares $= \sum_{i=1}^{n} \widehat{y}_i^2 = \sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2$ which can be obtained by subtracting $\bar{Y} = \overline{\widehat{Y}} = 6.5$ from the $\widehat{Y}_i$ column, squaring that column and summing. This yields 34.404762. This could have also been obtained as

$$\sum_{i=1}^{n} \widehat{y}_i^2 = \widehat{\beta}_{OLS}^2 \sum_{i=1}^{n} x_i^2 = (0.809524)^2(52.5) = 34.404762.$$

A final check is that $\sum_{i=1}^{n} \widehat{y}_i^2 = \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} e_i^2 = 36.9 - 2.495238 = 34.404762$ as required.

Recall, that $R^2 = r_{xy}^2 = (\sum_{i=1}^{n} x_i y_i)^2/(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i^2) = (42.5)^2/(52.5)(36.9) = 0.9324$. This could have also been obtained as $R^2 = 1 - (\sum_{i=1}^{n} e_i^2 / \sum_{i=1}^{n} y_i^2) = 1 - (2.495238/36.9) = 0.9324$, or as

$$R^2 = r_{y\hat{y}}^2 = \sum_{i=1}^{n} \widehat{y}_i^2 / \sum_{i=1}^{n} y_i^2 = 34.404762/36.9 = 0.9324.$$

This means that personal disposable income explains 93.24% of the variation in consumption. A plot of the actual, predicted and residual values versus time is given in Figure 3.8. This was done using EViews.

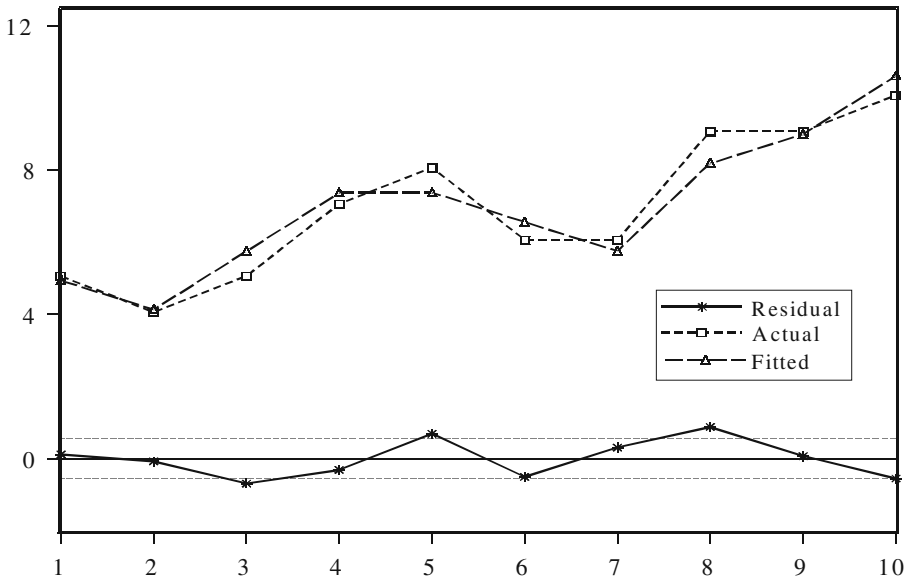


Figure 3.8 Residual Plot

3.10 Empirical Example

Table 3.2 gives (i) the logarithm of cigarette consumption (in packs) per person of smoking age (> 16 years) for 46 states in 1992, (ii) the logarithm of real price of cigarettes in each state, and (iii) the logarithm of real disposable income per capita in each state. This is drawn from Baltagi and Levin (1992) study on dynamic demand for cigarettes. It can be downloaded as Cigarett.dat from the Springer web site.

Table 3.2 Cigarette Consumption Data

LNC: log of consumption (in packs) per person of smoking age (>16)
LNP: log of real price (1983$/pack)
LNY: log of real disposable income per-capita (in thousand 1983$)

OBS	STATE	LNC	LNP	LNY
1	AL	4.96213	0.20487	4.64039
2	AZ	4.66312	0.16640	4.68389
3	AR	5.10709	0.23406	4.59435
4	CA	4.50449	0.36399	4.88147
5	CT	4.66983	0.32149	5.09472
6	DE	5.04705	0.21929	4.87087
7	DC	4.65637	0.28946	5.05960
8	FL	4.80081	0.28733	4.81155
9	GA	4.97974	0.12826	4.73299
10	ID	4.74902	0.17541	4.64307
11	IL	4.81445	0.24806	4.90387
12	IN	5.11129	0.08992	4.72916
13	IA	4.80857	0.24081	4.74211
14	KS	4.79263	0.21642	4.79613
15	KY	5.37906	–0.03260	4.64937
16	LA	4.98602	0.23856	4.61461
17	ME	4.98722	0.29106	4.75501
18	MD	4.77751	0.12575	4.94692
19	MA	4.73877	0.22613	4.99998
20	MI	4.94744	0.23067	4.80620
21	MN	4.69589	0.34297	4.81207
22	MS	4.93990	0.13638	4.52938
23	MO	5.06430	0.08731	4.78189
24	MT	4.73313	0.15303	4.70417
25	NE	4.77558	0.18907	4.79671
26	NV	4.96642	0.32304	4.83816
27	NH	5.10990	0.15852	5.00319
28	NJ	4.70633	0.30901	5.10268
29	NM	4.58107	0.16458	4.58202
30	NY	4.66496	0.34701	4.96075
31	ND	4.58237	0.18197	4.69163
32	OH	4.97952	0.12889	4.75875
33	OK	4.72720	0.19554	4.62730
34	PA	4.80363	0.22784	4.83516
35	RI	4.84693	0.30324	4.84670
36	SC	5.07801	0.07944	4.62549
37	SD	4.81545	0.13139	4.67747
38	TN	5.04939	0.15547	4.72525
39	TX	4.65398	0.28196	4.73437
40	UT	4.40859	0.19260	4.55586
41	VT	5.08799	0.18018	4.77578
42	VA	4.93065	0.11818	4.85490
43	WA	4.66134	0.35053	4.85645
44	WV	4.82454	0.12008	4.56859
45	WI	4.83026	0.22954	4.75826
46	WY	5.00087	0.10029	4.71169

Data: Cigarette Consumption of 46 States in 1992

Table 3.3 Cigarette Consumption Regression

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob > F
Model	1	0.48048	0.48048	18.084	0.0001
Error	44	1.16905	0.02657		
Root MSE		0.16300	R-square	0.2913	
Dep Mean		4.84784	Adj R-sq	0.2752	
C.V.		3.36234			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	5.094108	0.06269897	81.247	0.0001
LNP	1	–1.198316	0.28178857	–4.253	0.0001

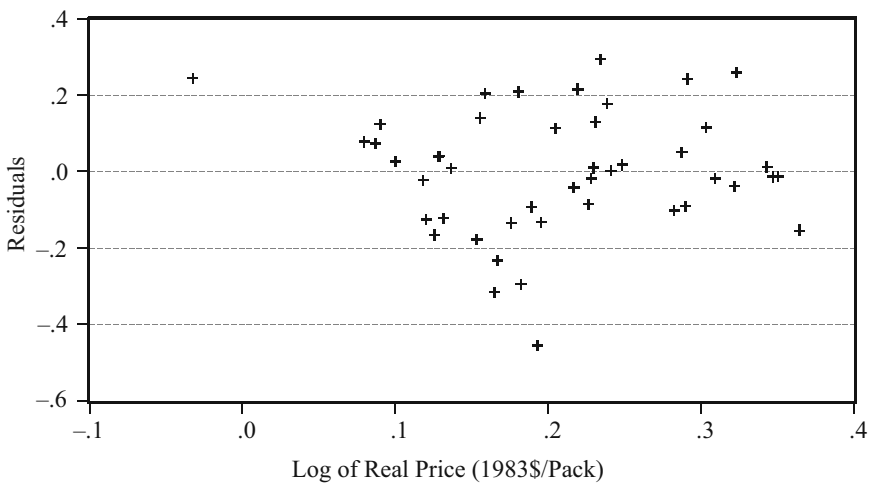


Figure 3.9 Residuals versus LNP

Table 3.3 gives the SAS output for the regression of log$C$ on log$P$. The price elasticity of demand for cigarettes in this simple model is ($d$log$C$/log$P$) which is the slope coefficient. This is estimated to be -1.198 with a standard error of 0.282. This says that a 10% increase in real price of cigarettes has an estimated 12% drop in per capita consumption of cigarettes. The R^2 of this regression is 0.29, s^2 is given by the Mean Square Error of the regression which is 0.0266. Figure 3.9 plots the residuals of this regression versus the independent variable, while Figure 3.10 plots the predictions along with the 95% confidence interval band for these predictions. One observation clearly stands out as an influential observation given its distance from the rest of the data and that is the observation for Kentucky, a producer state with very low real price. This observation almost anchors the straight line fit through the data. More on influential observations in Chapter 8.

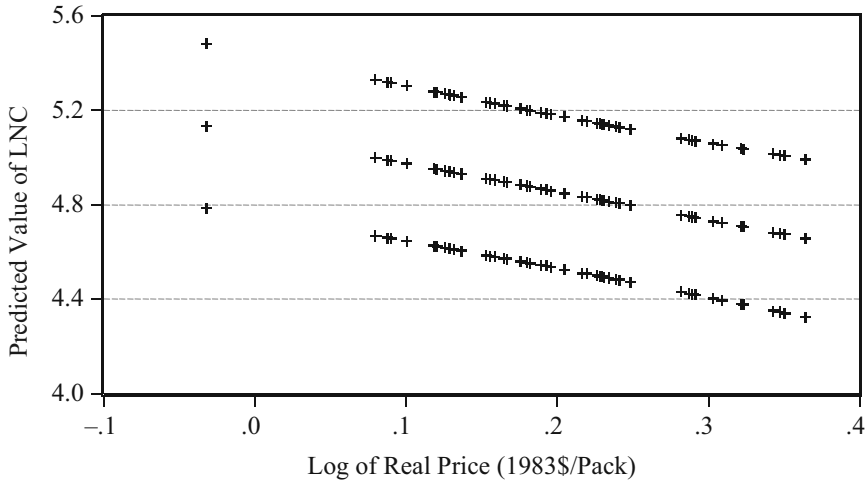


Figure 3.10 95% Confidence Band for Predicted Values

Problems

1. For the simple regression with a constant $Y_i = \alpha + \beta X_i + u_i$, given in equation (3.1) verify the following numerical properties of the OLS estimators:

$$\sum_{i=1}^{n} e_i = 0, \sum_{i=1}^{n} e_i X_i = 0, \sum_{i=1}^{n} e_i \widehat{Y}_i = 0, \sum_{i=1}^{n} \widehat{Y}_i = \sum_{i=1}^{n} Y_i$$

2. For the regression with only a constant $Y_i = \alpha + u_i$ with $u_i \sim \text{IID}(0, \sigma^2)$, show that the least squares estimate of $\widehat{\alpha}$ is $\widehat{\alpha}_{OLS} = \bar{Y}$, $\text{var}(\widehat{\alpha}_{OLS}) = \sigma^2/n$, and the residual sums of squares is $\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$.

3. For the simple regression *without* a constant $Y_i = \beta X_i + u_i$, with $u_i \sim \text{IID}(0, \sigma^2)$.

   (a) Derive the OLS estimator of β and find its variance.

   (b) What numerical properties of the OLS estimators described in problem 1 still hold for this model?

   (c) derive the maximum likelihood estimator of β and σ^2 under the assumption $u_i \sim \text{IIN}(0, \sigma^2)$.

   (d) Assume σ^2 is known. Derive the Wald, LM and LR tests for H_0; $\beta = 1$ versus H_1; $\beta \neq 1$.

4. Use the fact that $E(\sum_{i=1}^{n} x_i u_i)^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j E(u_i u_j)$; and assumptions 2 and 3 to prove equation (3.6).

5. Using the regression given in equation (3.1):

   (a) Show that $\widehat{\alpha}_{OLS} = \alpha + (\beta - \widehat{\beta}_{OLS})\bar{X} + \bar{u}$; and deduce that $E(\widehat{\alpha}_{OLS}) = \alpha$.

   (b) Using the fact that $\widehat{\beta}_{OLS} - \beta = \sum_{i=1}^{n} x_i u_i / \sum_{i=1}^{n} x_i^2$; use the results in part (a) to show that $\text{var}(\widehat{\alpha}_{OLS}) = \sigma^2[(1/n) + (\bar{X}^2 / \sum_{i=1}^{n} x_i^2)] = \sigma^2 \sum_{i=1}^{n} X_i^2 / n \sum_{i=1}^{n} x_i^2$.

   (c) Show that $\widehat{\alpha}_{OLS}$ is consistent for α.

   (d) Show that $\text{cov}(\widehat{\alpha}_{OLS}, \widehat{\beta}_{OLS}) = -\bar{X}\text{var}(\widehat{\beta}_{OLS}) = -\sigma^2 \bar{X} / \sum_{i=1}^{n} x_i^2$. This means that the sign of the covariance is determined by the sign of $\bar{X}$. If $\bar{X}$ is positive, this covariance will be negative. This also means that if $\widehat{\alpha}_{OLS}$ is over-estimated, $\widehat{\beta}_{OLS}$ will be under-estimated.

6. Using the regression given in equation (3.1):
   (a) Prove that $\widehat{\alpha}_{OLS} = \sum_{i=1}^{n} \lambda_i Y_i$ where $\lambda_i = (1/n) - \bar{X}w_i$ and $w_i = x_i / \sum_{i=1}^{n} x_i^2$.
   (b) Show that $\sum_{i=1}^{n} \lambda_i = 1$ and $\sum_{i=1}^{n} \lambda_i X_i = 0$.
   (c) Prove that any other linear estimator of α, say $\widetilde{\alpha} = \sum_{i=1}^{n} b_i Y_i$ must satisfy $\sum_{i=1}^{n} b_i = 1$ and $\sum_{i=1}^{n} b_i X_i = 0$ for $\widetilde{\alpha}$ to be unbiased for α.
   (d) Let $b_i = \lambda_i + f_i$; show that $\sum_{i=1}^{n} f_i = 0$ and $\sum_{i=1}^{n} f_i X_i = 0$.
   (e) Prove that $\text{var}(\widetilde{\alpha}) = \sigma^2 \sum_{i=1}^{n} b_i^2 = \sigma^2 \sum_{i=1}^{n} \lambda_i^2 + \sigma^2 \sum_{i=1}^{n} f_i^2 = \text{var}(\widehat{\alpha}_{OLS}) + \sigma^2 \sum_{i=1}^{n} f_i^2$.

7. (a) Differentiate (3.9) with respect to α and β and show that $\widehat{\alpha}_{MLE} = \widehat{\alpha}_{OLS}$, $\widehat{\beta}_{MLE} = \widehat{\beta}_{OLS}$.
   (b) Differentiate (3.9) with respect to σ^2 and show that $\widehat{\sigma}^2_{MLE} = \sum_{i=1}^{n} e_i^2/n$.

8. It is well known that a standard normal random variable $N(0,1)$ divided by a square root of a chi-squared random variable divided by its degrees of freedom $(\chi^2_\nu/\nu)^{\frac{1}{2}}$ results in a random variable that is t-distributed with ν degrees of freedom, provided the $N(0,1)$ and the χ^2 variables are independent, see Chapter 2. Use this fact to show that $(\widehat{\beta}_{OLS} - \beta)/[s/(\sum_{i=1}^{n} x_i^2)^{\frac{1}{2}}] \sim t_{n-2}$.

9. (a) Using the fact that $R^2 = \sum_{i=1}^{n} \widehat{y}_i^2 / \sum_{i=1}^{n} y_i^2$; $\widehat{y}_i = \widehat{\beta}_{OLS} x_i$; and $\widehat{\beta}_{OLS} = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2$, show that $R^2 = r^2_{xy}$ where,
   $$r^2_{xy} = (\sum_{i=1}^{n} x_i y_i)^2 / (\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i^2).$$
   (b) Using the fact that $y_i = \widehat{y}_i + e_i$, show that $\sum_{i=1}^{n} \widehat{y}_i y_i = \sum_{i=1}^{n} \widehat{y}_i^2$, and hence, deduce that $r^2_{y\widehat{y}} = (\sum_{i=1}^{n} y_i \widehat{y}_i)^2 / (\sum_{i=1}^{n} y_i^2)(\sum_{i=1}^{n} \widehat{y}_i^2)$ is equal to R^2.

10. *Prediction.* Consider the problem of predicting Y_0 from (3.11). Given X_0,
   (a) Show that $E(Y_0) = \alpha + \beta X_0$.
   (b) Show that $\widehat{Y}_0$ is unbiased for $E(Y_0)$.
   (c) Show that $\text{var}(\widehat{Y}_0) = \text{var}(\widehat{\alpha}_{OLS}) + X_0^2 \text{var}(\widehat{\beta}_{OLS}) + 2X_0 \text{cov}(\widehat{\alpha}_{OLS}, \widehat{\beta}_{OLS})$. Deduce that $\text{var}(\widehat{Y}_0) = \sigma^2[(1/n) + (X_0 - \bar{X})^2 / \sum_{i=1}^{n} x_i^2]$.
   (d) Consider a linear predictor of $E(Y_0)$, say $\widetilde{Y}_0 = \sum_{i=1}^{n} a_i Y_i$, show that $\sum_{i=1}^{n} a_i = 1$ and $\sum_{i=1}^{n} a_i X_i = X_0$ for this predictor to be unbiased for $E(Y_0)$.
   (e) Show that the $\text{var}(\widetilde{Y}_0) = \sigma^2 \sum_{i=1}^{n} a_i^2$. Minimize $\sum_{i=1}^{n} a_i^2$ subject to the restrictions given in (d). Prove that the resulting predictor is $\widetilde{Y}_0 = \widehat{\alpha}_{OLS} + \widehat{\beta}_{OLS} X_0$ and that the minimum variance is $\sigma^2[(1/n) + (X_0 - \bar{X})^2 / \sum_{i=1}^{n} x_i^2]$.

11. *Optimal Weighting of Unbiased Estimators.* This is based on Baltagi (1995). For the simple regression without a constant $Y_i = \beta X_i + u_i, i = 1, 2, \ldots, N$; where β is a scalar and $u_i \sim \text{IID}(0, \sigma^2)$ independent of X_i. Consider the following three unbiased estimators of β:
   $$\widehat{\beta}_1 = \sum_{i=1}^{n} X_i Y_i / \sum_{i=1}^{n} X_i^2, \; \widehat{\beta}_2 = \bar{Y}/\bar{X}$$
   and
   $$\widehat{\beta}_3 = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) / \sum_{i=1}^{n} (X_i - \bar{X})^2,$$
   where $\bar{X} = \sum_{i=1}^{n} X_i/n$ and $\bar{Y} = \sum_{i=1}^{n} Y_i/n$.
   (a) Show that $\text{cov}(\widehat{\beta}_1, \widehat{\beta}_2) = \text{var}(\widehat{\beta}_1) > 0$, and that $\rho_{12} = (\text{the correlation coefficient of } \widehat{\beta}_1 \text{ and } \widehat{\beta}_2) = [\text{var}(\widehat{\beta}_1)/\text{var}(\widehat{\beta}_2)]^{\frac{1}{2}}$ with $0 < \rho_{12} \leq 1$. Show that the optimal combination of $\widehat{\beta}_1$ and $\widehat{\beta}_2$, given by $\widehat{\beta} = \alpha\widehat{\beta}_1 + (1-\alpha)\widehat{\beta}_2$ where $-\infty < \alpha < \infty$ occurs at $\alpha^* = 1$. Optimality here refers to minimizing the variance. **Hint**: Read the paper by Samuel-Cahn (1994).

(b) Similarly, show that $\text{cov}(\widehat{\beta}_1, \widehat{\beta}_3) = \text{var}(\widehat{\beta}_1) > 0$, and that $\rho_{13} = (\text{the correlation coefficient of } \widehat{\beta}_1 \text{ and } \widehat{\beta}_3) = [\text{var}(\widehat{\beta}_1)/\text{var}(\widehat{\beta}_3)]^{\frac{1}{2}} = (1-\rho_{12}^2)^{\frac{1}{2}}$ with $0 < \rho_{13} < 1$. Conclude that the optimal combination $\widehat{\beta}_1$ and $\widehat{\beta}_3$ is again $\alpha^* = 1$.

(c) Show that $\text{cov}(\widehat{\beta}_2, \widehat{\beta}_3) = 0$ and that optimal combination of $\widehat{\beta}_2$ and $\widehat{\beta}_3$ is $\widehat{\beta} = (1-\rho_{12}^2)\widehat{\beta}_3 + \rho_{12}^2\widehat{\beta}_2 = \widehat{\beta}_1$. This exercise demonstrates a more general result, namely that the BLUE of β in this case $\widehat{\beta}_1$, has a positive correlation with any other linear unbiased estimator of β, and that this correlation can be easily computed from the ratio of the variances of these two estimators.

12. *Efficiency as Correlation.* This is based on Oksanen (1993). Let $\widehat{\beta}$ denote the Best Linear Unbiased Estimator of β and let $\widetilde{\beta}$ denote any linear unbiased estimator of β. Show that the relative efficiency of $\widetilde{\beta}$ with respect to $\widehat{\beta}$ is the squared correlation coefficient between $\widehat{\beta}$ and $\widetilde{\beta}$. **Hint**: Compute the variance of $\widehat{\beta} + \lambda(\widehat{\beta} - \widetilde{\beta})$ for any λ. This variance is minimized at $\lambda = 0$ since $\widehat{\beta}$ is BLUE. This should give you the result that $E(\widehat{\beta}^2) = E(\widehat{\beta}\widetilde{\beta})$ which in turn proves the required result, see Zheng (1994).

13. For the numerical illustration given in section 3.9, what happens to the least squares regression coefficient estimates $(\widehat{\alpha}_{OLS}, \widehat{\beta}_{OLS})$, s^2, the estimated $se(\widehat{\alpha}_{OLS})$ and $se(\widehat{\beta}_{OLS})$, t-statistic for $\widehat{\alpha}_{OLS}$ and $\widehat{\beta}_{OLS}$ for $H_0^a; \alpha = 0$, and $H_0^b; \beta = 0$ and R^2 when:

(a) Y_i is regressed on $X_i + 5$ rather than X_i. In other words, we add a constant 5 to each observation of the explanatory variable X_i and rerun the regression. It is very instructive to see how the computations in Table 3.1 are affected by this simple transformation on X_i.

(b) $Y_i + 2$ is regressed on X_i. In other words, a constant 2 is added to Y_i.

(c) Y_i is regressed on $2X_i$. (A constant 2 is multiplied by X_i).

14. For the cigarette consumption data given in Table 3.2.

(a) Give the descriptive statistics for log$C$, log$P$ and log$Y$. Plot their histogram. Also, plot log$C$ versus log$Y$ and log$C$ versus log$P$. Obtain the correlation matrix of these variables.

(b) Run the regression of log$C$ on log$Y$. What is the income elasticity estimate? What is its standard error? Test the null hypothesis that this elasticity is zero. What is the s and R^2 of this regression?

(c) Show that the square of the simple correlation coefficient between log$C$ and log$Y$ is equal to R^2. Show that the square of the correlation coefficient between the fitted and actual values of log$C$ is also equal to R^2.

(d) Plot the residuals versus income. Also, plot the fitted values along with their 95% confidence band.

15. Consider the simple regression with no constant: $Y_i = \beta X_i + u_i \quad i = 1, 2, \ldots, n$

where $u_i \sim \text{IID}(0, \sigma^2)$ independent of X_i. Theil (1971) showed that among all *linear* estimators in Y_i, the *minimum mean square* estimator for β, i.e., that which minimizes $E(\widetilde{\beta} - \beta)^2$ is given by

$$\widetilde{\beta} = \beta^2 \textstyle\sum_{i=1}^n X_i Y_i / (\beta^2 \sum_{i=1}^n X_i^2 + \sigma^2).$$

(a) Show that $E(\widetilde{\beta}) = \beta/(1+c)$, where $c = \sigma^2/\beta^2 \sum_{i=1}^n X_i^2 > 0$.

(b) Conclude that the Bias $(\widetilde{\beta}) = E(\widetilde{\beta}) - \beta = -[c/(1+c)]\beta$. Note that this bias is positive (negative) when β is negative (positive). This also means that $\widetilde{\beta}$ is biased towards zero.

(c) Show that $\text{MSE}(\widetilde{\beta}) = E(\widetilde{\beta} - \beta)^2 = \sigma^2/[\sum_{i=1}^n X_i^2 + (\sigma^2/\beta^2)]$. Conclude that it is smaller than the $\text{MSE}(\widehat{\beta}_{OLS})$.

Table 3.4 Energy Data for 20 countries

Country	RGDP (in 10^6 1975 U.S.$'s)	EN 10^6 Kilograms Coal Equivalents
Malta	1251	456
Iceland	1331	1124
Cyprus	2003	1211
Ireland	11788	11053
norway	27914	26086
Finland	28388	26405
Portugal	30642	12080
Denmark	34540	27049
Greece	38039	20119
Switzerland	42238	23234
Austria	45451	30633
Sweden	59350	45132
Belgium	62049	58894
Netherlands	82804	84416
Turkey	91946	32619
Spain	159602	88148
Italy	265863	192453
U.K.	279191	268056
France	358675	233907
W. Germany	428888	352677

16. Table 3.4 gives cross-section Data for 1980 on real gross domestic product (RGDP) and aggregate energy consumption (EN) for 20 countries

    (a) Enter the data and provide descriptive statistics. Plot the histograms for RGDP and EN. Plot EN versus RGDP.

    (b) Estimate the regression:

    $$\log(En) = \alpha + \beta \log(RGDP) + u.$$

    Be sure to plot the residuals. What do they show?

    (c) Test $H_0; \beta = 1$.

    (d) One of your Energy data observations has a misplaced decimal. Multiply it by 1000. Now repeat parts (a), (b) and (c).

    (e) Was there any reason for ordering the data from the lowest to highest energy consumption? Explain.

    Lesson Learned: Always plot the residuals. Always check your data very carefully.

17. Using the Energy Data given in Table 3.4, corrected as in problem 16 part (d), is it legitimate to reverse the form of the equation?

    $$log(RDGP) = \gamma + \delta log(En) + \epsilon$$

    (a) Economically, does this change the interpretation of the equation? Explain.

    (b) Estimate this equation and compare R^2 of this equation with that of the previous problem. Also, check if $\widehat{\delta} = 1/\widehat{\beta}$. Why are they different?

(c) Statistically, by reversing the equation, which assumptions do we violate?

(d) Show that $\widehat{\delta}\widehat{\beta} = R^2$.

(e) *Effects of changing units in which variables are measured*. Suppose you measured energy in BTU's instead of kilograms of coal equivalents so that the original series was multiplied by 60. How does it change α and β in the following equations?

$$\log(En) = \alpha + \beta \log(RDGP) + u \qquad En = \alpha^* + \beta^* RGDP + \nu$$

Can you explain why $\widehat{\alpha}$ changed, but not $\widehat{\beta}$ for the log-log model, whereas both $\widehat{\alpha}^*$and $\widehat{\beta}^*$ changed for the linear model?

(f) For the log-log specification and the linear specification, compare the GDP elasticity for Malta and W. Germany. Are both equally plausible?

(g) Plot the residuals from both linear and log-log models. What do you observe?

(h) Can you compare the R^2 and standard errors from both models in part (g)? **Hint**: Retrieve $\log(En)$ and $\widehat{\log}(\mathrm{E}n)$ in the log-log equation, exponentiate, then compute the residuals and s. These are comparable to those obtained from the linear model.

18. For the model considered in problem 16: $\log(En) = \alpha + \beta \log(RGDP) + u$ and measuring energy in BTU's (like part (e) of problem 17).

(a) What is the 95% confidence prediction interval at the sample mean?

(b) What is the 95% confidence prediction interval for Malta?

(c) What is the 95% confidence prediction interval for West Germany?

References

Additional readings on the material covered in this chapter can be found in:

Baltagi, B.H. (1995), "Optimal Weighting of Unbiased Estimators," *Econometric Theory*, Problem 95.3.1, 11:637.

Baltagi, B.H. and D. Levin (1992), "Cigarette Taxation: Raising Revenues and Reducing Consumption," *Structural Change and Economic Dynamics*, 3: 321-335.

Belsley, D.A., E. Kuh and R.E. Welsch (1980), *Regression Diagnostics* (Wiley: New York).

Greene, W. (1993), *Econometric Analysis* (Macmillian: New York).

Gujarati, D. (1995), *Basic Econometrics* (McGraw-Hill: New York).

Johnston, J. (1984), *Econometric Methods* (McGraw-Hill: New York).

Kelejian, H. and W. Oates (1989), *Introduction to Econometrics* (Harper and Row: New York).

Kennedy, P. (1992), *A Guide to Econometrics* (MIT Press: Cambridge).

Kmenta, J. (1986), *Elements of Econometrics* (Macmillan: New York).

Maddala, G.S. (1992), *Introduction to Econometrics* (Macmillan: New York).

Oksanen, E.H. (1993), "Efficiency as Correlation," *Econometric Theory*, Problem 93.1.3, 9: 146.

Samuel-Cahn, E. (1994), "Combining Unbiased Estimators," *The American Statistician*, 48: 34-36.

Wallace, D. and L. Silver (1988), *Econometrics: An Introduction* (Addison Wesley: New York).

Zheng, J.X. (1994), "Efficiency as Correlation," *Econometric Theory*, Solution 93.1.3, 10: 228.

Appendix
Centered and Uncentered R^2

From the OLS regression on (3.1) we get

$$Y_i = \widehat{Y}_i + e_i \quad i = 1, 2, \ldots, n \tag{A.1}$$

where $\widehat{Y}_i = \widehat{\alpha}_{OLS} + X_i\widehat{\beta}_{OLS}$. Squaring and summing the above equation we get

$$\sum_{i=1}^n Y_i^2 = \sum_{i=1}^n \widehat{Y}_i^2 + \sum_{i=1}^n e_i^2 \tag{A.2}$$

since $\sum_{i=1}^n \widehat{Y}_i e_i = 0$. The *uncentered* R^2 is given by

$$\text{uncentered } R^2 = 1 - \sum_{i=1}^n e_i^2 / \sum_{i=1}^n Y_i^2 = \sum_{i=1}^n \widehat{Y}_i^2 / \sum_{i=1} Y_i^2 \tag{A.3}$$

Note that the total sum of squares for Y_i is *not* expressed in deviation from the sample mean $\bar{Y}$. In other words, the uncentered R^2 is the proportion of variation of $\sum_{i=1}^n Y_i^2$ that is explained by the regression on X. Regression packages usually report the *centered* R^2 which was defined in section 3.6 as $1 - (\sum_{i=1}^n e_i^2 / \sum_{i=1}^n y_i^2)$ where $y_i = Y_i - \bar{Y}$. The latter measure focuses on explaining the variation in Y_i *after* fitting the constant.

From section 3.6, we have seen that a naive model with only a constant in it gives $\bar{Y}$ as the estimate of the constant, see also problem 2. The variation in Y_i that is not explained by this naive model is $\sum_{i=1}^n y_i^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2$. Subtracting $n\bar{Y}^2$ from both sides of (A.2) we get

$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^n \widehat{Y}_i^2 - n\bar{Y}^2 + \sum_{i=1}^n e_i^2$$

and the centered R^2 is

$$\text{centered } R^2 = 1 - (\sum_{i=1}^n e_i^2 / \sum_{i=1}^n y_i^2) = (\sum_{i=1}^n \widehat{Y}_i^2 - n\bar{Y}^2) / \sum_{i=1}^n y_i^2 \tag{A.4}$$

If there is a constant in the model $\bar{Y} = \overline{\widehat{Y}}$, see section 3.6, and $\sum_{i=1}^n \widehat{y}_i^2 = \sum_{i=1}^n (\widehat{Y}_i - \overline{\widehat{Y}})^2 = \sum_{i=1}^n \widehat{Y}_i^2 - n\bar{Y}^2$. Therefore, the centered $R^2 = \sum_{i=1}^n \widehat{y}_i^2 / \sum_{i=1}^n y_i^2$ which is the R^2 reported by regression packages. If there is no constant in the model, some regression packages give you the option of (no constant) and the R^2 reported is usually the uncentered R^2. Check your regression package documentation to verify what you are getting. We will encounter uncentered R^2 again in constructing test statistics using regressions, see for example Chapter 11.

CHAPTER 4
Multiple Regression Analysis

4.1 Introduction

So far we have considered only one regressor X besides the constant in the regression equation. Economic relationships usually include more than one regressor. For example, a demand equation for a product will usually include real price of that product in addition to real income as well as real price of a competitive product and the advertising expenditures on this product. In this case

$$Y_i = \alpha + \beta_2 X_{2i} + \beta_3 X_{3i} + .. + \beta_K X_{Ki} + u_i \quad i = 1, 2, \ldots, n \tag{4.1}$$

where Y_i denotes the i-th observation on the dependent variable Y, in this case the sales of this product. X_{ki} denotes the i-th observation on the independent variable X_k for $k = 2, \ldots, K$ in this case, own price, the competitor's price and advertising expenditures. α is the intercept and $\beta_2, \beta_3, \ldots, \beta_K$ are the $(K-1)$ slope coefficients. The u_i's satisfy the classical assumptions 1-4 given in Chapter 3. Assumption 4 is modified to include all the X's appearing in the regression, i.e., every X_k for $k = 2, \ldots, K$, is uncorrelated with the u_i's with the property that $\sum_{i=1}^{n}(X_{ki} - \bar{X}_k)^2/n$ where $\bar{X}_k = \sum_{i=1}^{n} X_{ki}/n$ has a finite probability limit which is different from zero.

Section 4.2 derives the OLS normal equations of this multiple regression model and discovers that an additional assumption is needed for these equations to yield a unique solution.

4.2 Least Squares Estimation

As explained in Chapter 3, least squares minimizes the residual sum of squares where the residuals are now given by $e_i = Y_i - \widehat{\alpha} - \sum_{k=2}^{K} \widehat{\beta}_k X_{ki}$ and $\widehat{\alpha}$ and $\widehat{\beta}_k$ denote guesses on the regression parameters α and β_k, respectively. The residual sum of squares

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{\alpha} - \widehat{\beta}_2 X_{2i} - .. - \widehat{\beta}_K X_{Ki})^2$$

is minimized by the following K first-order conditions:

$$\begin{aligned} &\partial(\textstyle\sum_{i=1}^{n} e_i^2)/\partial\widehat{\alpha} = -2\sum_{i=1}^{n} e_i = 0 \\ &\partial(\textstyle\sum_{i=1}^{n} e_i^2)/\partial\widehat{\beta}_k = -2\sum_{i=1}^{n} e_i X_{ki} = 0, \text{ for } k = 2, \ldots, K. \end{aligned} \tag{4.2}$$

or, equivalently

$$\begin{aligned} &\textstyle\sum_{i=1}^{n} Y_i = \widehat{\alpha} n + \widehat{\beta}_2 \sum_{i=1}^{n} X_{2i} + .. + \widehat{\beta}_K \sum_{i=1}^{n} X_{ki} \\ &\textstyle\sum_{i=1}^{n} Y_i X_{2i} = \widehat{\alpha} \sum_{i=1}^{n} X_{2i} + \widehat{\beta}_2 \sum_{i=1}^{n} X_{2i}^2 + .. + \widehat{\beta}_K \sum_{i=1}^{n} X_{2i} X_{Ki} \\ &\quad\vdots \qquad\qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad\qquad \vdots \\ &\textstyle\sum_{i=1}^{n} Y_i X_{Ki} = \widehat{\alpha} \sum_{i=1}^{n} X_{Ki} + \widehat{\beta}_2 \sum_{i=1}^{n} X_{2i} X_{Ki} + .. + \widehat{\beta}_K \sum_{i=1}^{n} X_{Ki}^2 \end{aligned} \tag{4.3}$$

where the first equation multiplies the regression equation by the constant and sums, the second equation multiplies the regression equation by X_2 and sums, and the K-th equation multiplies the regression equation by X_K and sums. $\sum_{i=1}^{n} u_i = 0$ and $\sum_{i=1}^{n} u_i X_{ki} = 0$ for $k = 2, \ldots, K$ are implicitly imposed to arrive at (4.3). Solving these K equations in K unknowns, we get the OLS estimators. This can be done more succinctly in matrix form, see Chapter 7. Assumptions 1-4 insure that the OLS estimator is BLUE. Assumption 5 introduces normality and as a result the OLS estimator is also (i) a maximum likelihood estimator, (ii) it is normally distributed, and (iii) it is minimum variance unbiased. Normality also allows test of hypotheses. Without the normality assumption, one has to appeal to the Central Limit Theorem and the fact that the sample is large to perform hypotheses testing.

In order to make sure we can solve for the OLS estimators in (4.3) we need to impose one further assumption on the model besides those considered in Chapter 3.

Assumption 6: *No perfect multicollinearity*, i.e., the explanatory variables are not perfectly correlated with each other. This assumption states that, no explanatory variable X_k for $k = 2, \ldots, K$ is a perfect *linear* combination of the other X's. If assumption 6 is violated, then one of the equations in (4.2) or (4.3) becomes redundant and we would have $K - 1$ linearly independent equations in K unknowns. This means that we cannot solve uniquely for the OLS estimators of the K coefficients.

Example 1: If $X_{2i} = 3X_{4i} - 2X_{5i} + X_{7i}$ for $i = 1, \ldots, n$, then multiplying this relationship by e_i and summing over i we get

$$\sum_{i=1}^{n} X_{2i} e_i = 3 \sum_{i=1}^{n} X_{4i} e_i - 2 \sum_{i=1}^{n} X_{5i} e_i + \sum_{i=1}^{n} X_{7i} e_i.$$

This means that the second OLS normal equation in (4.2) can be represented as a perfect linear combination of the fourth, fifth and seventh OLS normal equations. Knowing the latter three equations, the second equation adds no new information. Alternatively, one could substitute this relationship in the original regression equation (4.1). After some algebra, X_2 would be eliminated and the resulting equation becomes:

$$\begin{aligned} Y_i &= \alpha + \beta_3 X_{3i} + (3\beta_2 + \beta_4) X_{4i} + (\beta_5 - 2\beta_2) X_{5i} + \beta_6 X_{6i} + (\beta_2 + \beta_7) X_{7i} \\ &\quad + .. + \beta_K X_{Ki} + u_i. \end{aligned} \tag{4.4}$$

Note that the coefficients of X_{4i}, X_{5i} and X_{7i} are now $(3\beta_2 + \beta_4)$, $(\beta_5 - 2\beta_2)$ and $(\beta_2 + \beta_7)$, respectively. All of which are contaminated by β_2. These linear combinations of β_2, β_4, β_5 and β_7 can be estimated from regression (4.4) which excludes X_{2i}. In fact, the other X's, not contaminated by this perfect linear relationship, will have coefficients that are not contaminated by β_2 and hence are themselves estimable using OLS. However, β_2, β_4, β_5 and β_7 cannot be estimated separately. Perfect multicollinearity means that we cannot separate the influence on Y of the independent variables that are perfectly related. Hence, assumption 6 of no perfect multicollinearity is needed to guarantee a unique solution of the OLS normal equations. Note that it applies to perfect linear relationships and does *not* apply to perfect non-linear relationships among the independent variables. In other words, one can include X_{1i} and X_{1i}^2 like (years of experience) and (years of experience)^2 in an equation explaining earnings of individuals. Although, there is a perfect quadratic relationship between these independent variables, this is not a perfect *linear* relationship and therefore, does not cause perfect multicollinearity.

4.3 Residual Interpretation of Multiple Regression Estimates

Although we did not derive an explicit solution for the OLS estimators of the β's, we know that they are the solutions to (4.2) or (4.3). Let us focus on one of these estimators, say $\widehat{\beta}_2$, the OLS estimator of β_2, the partial derivative of Y_i with respect to X_{2i}. As a solution to (4.2) or (4.3), $\widehat{\beta}_2$ is a multiple regression coefficient estimate of β_2. Alternatively, we can interpret $\widehat{\beta}_2$ as a simple linear regression coefficient.

Claim 1: (i) Run the regression of X_2 on all the *other* X's in (4.1), and obtain the residuals $\widehat{\nu}_2$, i.e., $X_2 = \widehat{X}_2 + \widehat{\nu}_2$. (ii) Run the simple regression of Y on $\widehat{\nu}_2$, the resulting estimate of the slope coefficient is $\widehat{\beta}_2$.

The first regression essentially cleans out the effect of the other X's from X_2, leaving the variation unique to X_2 in $\widehat{\nu}_2$. Claim 1 states that $\widehat{\beta}_2$ can be interpreted as a simple linear regression coefficient of Y on this residual. This is in line with the partial derivative interpretation of β_2. The proof of claim 1 is given in the Appendix. Using the results of the simple regression given in (3.4) with the regressor X_i replaced by the residual $\widehat{\nu}_2$, we get

$$\widehat{\beta}_2 = \textstyle\sum_{i=1}^n \widehat{\nu}_2 Y_i / \sum_{i=1}^n \widehat{\nu}_{2i}^2 \tag{4.5}$$

and from (3.6) we get

$$\text{var}(\widehat{\beta}_2) = \sigma^2 / \textstyle\sum_{i=1}^n \widehat{\nu}_{2i}^2 \tag{4.6}$$

An alternative interpretation of $\widehat{\beta}_2$ as a simple regression coefficient is the following:

Claim 2: (i) Run Y on all the *other* X's and get the predicted $\widetilde{Y}$ and the residuals, say $\widetilde{\omega}$. (ii) Run the simple linear regression of $\widetilde{\omega}$ on $\widehat{\nu}_2$. $\widehat{\beta}_2$ is the resulting estimate of the slope coefficient.

This regression cleans both Y and X_2 from the effect of the other X's and then regresses the cleaned out residuals of Y on those of X_2. Once again this is in line with the partial derivative interpretation of β_2. The proof of claim 2 is simple and is given in the Appendix.

These two interpretations of $\widehat{\beta}_2$ are important in that they provide an easy way of looking at a multiple regression in the context of a simple linear regression. Also, it says that there is no need to clean the effects of one X from the other X's to find its unique effect on Y. All one has to do is to include all these X's in the same multiple regression. Problem 1 verifies this result with an empirical example. This will also be proved using matrix algebra in Chapter 7.

Recall that $R^2 = 1 - RSS/TSS$ for any regression. Let R_2^2 be the R^2 for the regression of X_2 on all the other X's, then $R_2^2 = 1 - \sum_{i=1}^n \widehat{\nu}_{2i}^2 / \sum_{i=1}^n x_{2i}^2$ where $x_{2i} = X_{2i} - \bar{X}_2$ and $\bar{X}_2 = \sum_{i=1}^n X_{2i}/n$; $TSS = \sum_{i=1}^n (X_{2i} - \bar{X}_2)^2 = \sum_{i=1}^n x_{2i}^2$ and $RSS = \sum_{i=1}^n \widehat{\nu}_{2i}^2$. Equivalently, $\sum_{i=1}^n \widehat{\nu}_{2i}^2 = \sum_{i=1}^n x_{2i}^2 (1 - R_2^2)$ and the

$$\text{var}(\widehat{\beta}_2) = \sigma^2 / \textstyle\sum_{i=1}^n \widehat{\nu}_{2i}^2 = \sigma^2 / \sum_{i=1}^n x_{2i}^2 (1 - R_2^2) \tag{4.7}$$

This means that the larger R_2^2, the smaller is $(1 - R_2^2)$ and the larger is $\text{var}(\widehat{\beta}_2)$ holding σ^2 and $\sum_{i=1}^n x_{2i}^2$ fixed. This shows the relationship between multicollinearity and the variance of the OLS estimates. High multicollinearity between X_2 and the other X's will result in high R_2^2 which in turn implies high variance for $\widehat{\beta}_2$. Perfect multicollinearity is the extreme case where $R_2^2 = 1$. This in turn implies an infinite variance for $\widehat{\beta}_2$. In general, high multicollinearity among the regressors yields imprecise estimates for these highly correlated variables. The least

squares regression estimates are still unbiased as long as assumptions 1 and 4 are satisfied, but these estimates are unreliable as reflected by their high variances. However, it is important to note that a low σ^2 and a high $\sum_{i=1}^{n} x_{2i}^2$ could counteract the effect of a high R_2^2 leading to a significant t-statistic for $\widehat{\beta}_2$. Maddala (2001) argues that high intercorrelation among the explanatory variables are neither necessary nor sufficient to cause the multicollinearity problem. In practice, multicollinearity is sensitive to the addition or deletion of observations. More on this in Chapter 8. Looking at high intercorrelations among the explanatory variables is useful only as a complaint. It is more important to look at the standard errors and t-statistics to assess the seriousness of multicollinearity.

Much has been written on possible solutions to the multicollinearity problem, see Hill and Adkins (2001) for a recent summary. Credible candidates include: (i) obtaining *new and better data*, but this is rarely available; (ii) introducing *nonsample information* about the model parameters based on previous empirical research or economic theory. The problem with the latter solution is that we never truly know whether the information we introduce is good enough to reduce estimator Mean Square Error.

4.4 Overspecification and Underspecification of the Regression Equation

So far we have assumed that the true linear regression relationship is always correctly specified. This is likely to be violated in practice. In order to keep things simple, we consider the case where the true model is a simple regression with one regressor X_1.

True model: $Y_i = \alpha + \beta_1 X_{1i} + u_i$

with $u_i \sim \text{IID}(0, \sigma^2)$, but the estimated model is overspecified with the inclusion of an additional irrelevant variable X_2, i.e.,

Estimated model: $\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i}$

From the previous section, it is clear that $\widehat{\beta}_1 = \sum_{i=1}^{n} \widehat{\nu}_{1i} Y_i / \sum_{i=1}^{n} \widehat{\nu}_{1i}^2$ where $\widehat{\nu}_1$ is the OLS residuals of X_1 on X_2. Substituting the true model for Y we get

$$\widehat{\beta}_1 = \beta_1 \sum_{i=1}^{n} \widehat{\nu}_{1i} X_{1i} / \sum_{i=1}^{n} \widehat{\nu}_{1i}^2 + \sum_{i=1}^{n} \widehat{\nu}_{1i} u_i / \sum_{i=1}^{n} \widehat{\nu}_{1i}^2$$

since $\sum_{i=1}^{n} \widehat{\nu}_{1i} = 0$. But, $X_{1i} = \widehat{X}_{1i} + \widehat{\nu}_{1i}$ and $\sum_{i=1}^{n} \widehat{X}_{1i} \widehat{\nu}_{1i} = 0$ implying that $\sum_{i=1}^{n} \widehat{\nu}_{1i} X_{1i} = \sum_{i=1}^{n} \widehat{\nu}_{1i}^2$. Hence,

$$\widehat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} \widehat{\nu}_{1i} u_i / \sum_{i=1}^{n} \widehat{\nu}_{1i}^2 \tag{4.8}$$

and $E(\widehat{\beta}_1) = \beta_1$ since $\widehat{\nu}_1$ is a linear combination of the X's, and $E(X_k u) = 0$ for $k = 1, 2$. Also,

$$\text{var}(\widehat{\beta}_1) = \sigma^2 / \sum_{i=1}^{n} \widehat{\nu}_{1i}^2 = \sigma^2 / \sum_{i=1}^{n} x_{1i}^2 (1 - R_1^2) \tag{4.9}$$

where $x_{1i} = X_{1i} - \bar{X}_1$ and R_1^2 is the R^2 of the regression of X_1 on X_2. Using the true model to estimate β_1, one would get $b_1 = \sum_{i=1}^{n} x_{1i} y_i / \sum_{i=1}^{n} x_{1i}^2$ with $E(b_1) = \beta_1$ and $\text{var}(b_1) =$

$\sigma^2/\sum_{i=1}^n x_{1i}^2$. Hence, $\text{var}(\widehat{\beta}_1) \geq \text{var}(b_1)$. Note also that in the overspecified model, the estimate for β_2 which has a true value of zero is given by

$$\widehat{\beta}_2 = \sum_{i=1}^n \widehat{\nu}_{2i} Y_i / \sum_{i=1}^n \widehat{\nu}_{2i}^2 \tag{4.10}$$

where $\widehat{\nu}_2$ is the OLS residual of X_2 on X_1. Substituting the true model for Y we get

$$\widehat{\beta}_2 = \sum_{i=1}^n \widehat{\nu}_{2i} u_i / \sum_{i=1}^n \widehat{\nu}_{2i}^2 \tag{4.11}$$

since $\sum_{i=1}^n \widehat{\nu}_{2i} X_{1i} = 0$ and $\sum_{i=1}^n \widehat{\nu}_{2i} = 0$. Hence, $E(\widehat{\beta}_2) = 0$ since $\widehat{\nu}_2$ is a linear combination of the X's and $E(X_k u) = 0$ for $k = 1, 2$. In summary, overspecification still yields unbiased estimates of β_1 and β_2, but the price is a higher variance.

Similarly, the true model could be a two-regressors model

$$\text{True model: } Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

where $u_i \sim \text{IID}(0, \sigma^2)$ but the estimated model is

$$\text{Estimated model: } \widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 X_{1i}$$

The estimated model omits a relevant variable X_2 and underspecifies the true relationship. In this case

$$\widehat{\beta}_1 = \sum_{i=1}^n x_{1i} Y_i / \sum_{i=1}^n x_{1i}^2 \tag{4.12}$$

where $x_{1i} = X_{1i} - \bar{X}_1$. Substituting the true model for Y we get

$$\widehat{\beta}_1 = \beta_1 + \beta_2 \sum_{i=1}^n x_{1i} X_{2i} / \sum_{i=1}^n x_{1i}^2 + \sum_{i=1}^n x_{1i} u_i / \sum_{i=1}^n x_{1i}^2 \tag{4.13}$$

Hence, $E(\widehat{\beta}_1) = \beta_1 + \beta_2 b_{12}$ since $E(x_1 u) = 0$ with $b_{12} = \sum_{i=1}^n x_{1i} X_{2i} / \sum_{i=1}^n x_{1i}^2$. Note that b_{12} is the regression slope estimate obtained by regressing X_2 on X_1 and a constant. Also, the

$$\text{var}(\widehat{\beta}_1) = E(\widehat{\beta}_1 - E(\widehat{\beta}_1))^2 = E(\sum_{i=1}^n x_{1i} u_i / \sum_{i=1}^n x_{1i}^2)^2 = \sigma^2 / \sum_{i=1}^n x_{1i}^2$$

which understates the variance of the estimate of β_1 obtained from the true model, i.e., $b_1 = \sum_{i=1}^n \widehat{\nu}_{1i} Y_i / \sum_{i=1}^n \widehat{\nu}_{1i}^2$ with

$$\text{var}(b_1) = \sigma^2 / \sum_{i=1}^n \widehat{\nu}_{1i}^2 = \sigma^2 / \sum_{i=1}^n x_{1i}^2 (1 - R_1^2) \geq \text{var}(\widehat{\beta}_1). \tag{4.14}$$

In summary, underspecification yields biased estimates of the regression coefficients and understates the variance of these estimates. This is also an example of imposing a zero restriction on β_2 when in fact it is not true. This introduces bias, because the restriction is wrong, but reduces the variance because it imposes more information even if this information may be false. We will encounter this general principle again when we discuss distributed lags in Chapter 6.

4.5 R-Squared versus R-Bar-Squared

Since OLS minimizes the residual sums of squares, adding one or more variables to the regression cannot increase this residual sums of squares. After all, we are minimizing over a larger dimension parameter set and the minimum there is smaller or equal to that over a subset of the parameter space, see problem 4. Therefore, for the same dependent variable Y, adding more variables makes $\sum_{i=1}^{n} e_i^2$ non-increasing and R^2 non-decreasing, since $R^2 = 1 - (\sum_{i=1}^{n} e_i^2 / \sum_{i=1}^{n} y_i^2)$. Hence, a criteria of selecting a regression that "maximizes R^2" does not make sense, since we can add more variables to this regression and improve on this R^2 (or at worst leave it the same). In order to penalize the researcher for adding an extra variable, one computes

$$\bar{R}^2 = 1 - [\textstyle\sum_{i=1}^{n} e_i^2/(n-K)]/[\sum_{i=1}^{n} y_i^2/(n-1)] \tag{4.15}$$

where $\sum_{i=1}^{n} e_i^2$ and $\sum_{i=1}^{n} y_i^2$ have been adjusted by their degrees of freedom. Note that the numerator is the s^2 of the regression and is equal to $\sum_{i=1}^{n} e_i^2/(n-K)$. This differs from the s^2 in Chapter 3 in the degrees of freedom. Here, it is $n-K$, because we have estimated K coefficients, or because (4.2) represents K relationships among the residuals. Therefore knowing $(n-K)$ residuals we can deduce the other K residuals from (4.2). $\sum_{i=1}^{n} e_i^2$ is non-increasing as we add more variables, but the degrees of freedom decrease by one with every added variable. Therefore, s^2 will decrease only if the effect of the $\sum_{i=1}^{n} e_i^2$ decrease outweighs the effect of the one degree of freedom loss on s^2. This is exactly the idea behind $\bar{R}^2$, i.e., penalizing each added variable by decreasing the degrees of freedom by one. Hence, this variable will increase $\bar{R}^2$ only if the reduction in $\sum_{i=1}^{n} e_i^2$ outweighs this loss, i.e., only if s^2 is decreased. Using the definition of $\bar{R}^2$, one can relate it to R^2 as follows:

$$(1-\bar{R}^2) = (1-R^2)[(n-1)/(n-K)] \tag{4.16}$$

4.6 Testing Linear Restrictions

In the simple linear regression chapter, we proved that the OLS estimates are BLUE provided assumptions 1 to 4 were satisfied. Then we imposed normality on the disturbances, assumption 5, and proved that the OLS estimators are in fact the maximum likelihood estimators. Then we derived the Cramér-Rao lower bound, and proved that these estimates are efficient. This will be done in matrix form in Chapter 7 for the multiple regression case. Under normality one can test hypotheses about the regression. Basically, any regression package will report the OLS estimates, their standard errors and the corresponding t-statistics for the null hypothesis that each individual coefficient is zero. These are tests of significance for each coefficient separately. But one may be interested in a joint test of significance for two or more coefficients simultaneously, or simply testing whether linear restrictions on the coefficients of the regression are satisfied. This will be developed more formally in Chapter 7. For now, all we assume is that the reader can perform regressions using his or her favorite software like EViews, STATA, SAS, TSP, SHAZAM, LIMDEP or GAUSS. The solutions to (4.2) or (4.3) result in the OLS estimates. These multiple regression coefficient estimates can be interpreted as simple regression estimates as shown in section 4.3. This allows a simple derivation of their standard errors. Now, we would like to use these regressions to test linear restrictions. The strategy followed is to impose these restrictions on the model and run the resulting restricted regression. The

corresponding Restricted Residual Sums of Squares is denoted by RRSS. Next, one runs the regression without imposing these linear restrictions to obtain the Unrestricted Residual Sums of Squares, which we denote by URSS. Finally, one forms the following F-statistic:

$$F = \frac{(RRSS - URSS)/\ell}{URSS/(n-K)} \sim F_{\ell,n-K} \tag{4.17}$$

where ℓ denotes the number of restrictions, and $n-K$ gives the degrees of freedom of the unrestricted model. The idea behind this test is intuitive. If the restrictions are true, then the RRSS should not be much different from the URSS. If RRSS is different from URSS, then we reject these restrictions. The denominator of the F-statistic is a consistent estimate of the unrestricted regression variance. Dividing by the latter makes the F-statistic invariant to units of measurement. Let us consider two examples:

Example 2: Testing the joint significance of two regression coefficients. For e.g., let us test the following null hypothesis $H_0; \beta_2 = \beta_3 = 0$. These are two restrictions $\beta_2 = 0$ and $\beta_3 = 0$ and they are to be tested jointly. We know how to test for $\beta_2 = 0$ alone or $\beta_3 = 0$ alone with individual t-tests. This is a test of joint significance of the two coefficients. Imposing this restriction, means the removal of X_2 and X_3 from the regression, i.e., running the regression of Y on $X_4, \ldots, X_K$ excluding X_2 and X_3. Hence, the number of parameters to be estimated becomes $(K-2)$ and the degrees of freedom of this restricted regression are $n-(K-2)$. The unrestricted regression is the one including all the X's in the model. Its degrees of freedom are $(n-K)$. The number of restrictions are 2 and this can also be inferred from the difference between the degrees of freedom of the restricted and unrestricted regressions. All the ingredients are now available for computing F in (4.17) and this will be distributed as $F_{2,n-K}$.

Example 3: Test the equality of two regression coefficients $H_0; \beta_3 = \beta_4$ against the alternative that $H_1; \beta_3 \neq \beta_4$. Note that H_0 can be rewritten as $H_0; \beta_3 - \beta_4 = 0$. This can be tested using a t-statistic that tests whether $d = \beta_3 - \beta_4$ is equal to zero. From the unrestricted regression, we can obtain $\widehat{d} = \widehat{\beta}_3 - \widehat{\beta}_4$ with $\text{var}(\widehat{d}) = \text{var}(\widehat{\beta}_3) + \text{var}(\widehat{\beta}_4) - 2\text{cov}(\widehat{\beta}_3, \widehat{\beta}_4)$. The variance-covariance matrix of the regression coefficients can be printed out with any regression package. In section 4.3, we gave these variances and covariances a simple regression interpretation. This means that $se(\widehat{d}) = \sqrt{\text{var}(\widehat{d})}$ and the t-statistic is simply $t = (\widehat{d} - 0)/se(\widehat{d})$ which is distributed as t_{n-K} under H_0. Alternatively, one can run an F-test with the RRSS obtained from running the following regression

$$Y_i = \alpha + \beta_2 X_{2i} + \beta_{3i}(X_{3i} + X_{4i}) + \beta_5 X_{5i} + .. + \beta_K X_{Ki} + u_i$$

with $\beta_3 = \beta_4$ substituted in for β_4. This regression has the variable $(X_{3i} + X_{4i})$ rather than X_{3i} and X_{4i} separately. The URSS is the regression of Y on all the X's in the model. The degrees of freedom of the resulting F-statistic are 1 and $n-K$. The numerator degree of freedom states that there is only one restriction. It will be proved in Chapter 7 that the square of the t-statistic is exactly equal to the F-statistic just derived. Both methods of testing are equivalent. The first one computes only the unrestricted regression and involves some further variance computations, while the latter involves running two regressions and computing the usual F-statistic.

Example 4: Test the joint hypothesis $H_0; \beta_3 = 1$ and $\beta_2 - 2\beta_4 = 0$. These two restrictions are usually obtained from prior information or imposed by theory. The first restriction is $\beta_3 = 1$.

The value 1 could have been any other constant. The second restriction shows that a linear combination of β_2 and β_4 is equal to zero. Substituting these restrictions in (4.1) we get

$$Y_i = \alpha + \beta_2 X_{2i} + X_{3i} + \tfrac{1}{2}\beta_2 X_{4i} + \beta_5 X_{5i} + .. + \beta_K X_{Ki} + u_i$$

which can be written as

$$Y_i - X_{3i} = \alpha + \beta_2 (X_{2i} + \tfrac{1}{2} X_{4i}) + \beta_5 X_{5i} + .. + \beta_K X_{Ki} + u_i$$

Therefore, the RRSS can be obtained by regressing $(Y - X_3)$ on $(X_2 + \frac{1}{2}X_4), X_5, \ldots, X_K$. This regression has $n - (K - 2)$ degrees of freedom. The URSS is the regression with all the X's included. The resulting F-statistic has 2 and $n - K$ degrees of freedom.

Example 5: Testing constant returns to scale in a Cobb-Douglas production function. $Q = AK^{\alpha}L^{\beta}E^{\gamma}M^{\delta}e^{u}$ is a Cobb-Douglas production function with capital(K), labor(L), energy(E) and material(M). Constant returns to scale means that a proportional increase in the inputs produces the same proportional increase in output. Let this proportional increase be λ, then $K^* = \lambda K$, $L^* = \lambda L$, $E^* = \lambda E$ and $M^* = \lambda M$. $Q^* = \lambda^{(\alpha+\beta+\gamma+\delta)} AK^{\alpha}L^{\beta}E^{\gamma}M^{\delta}e^{u} = \lambda^{(\alpha+\beta+\gamma+\delta)}Q$. For this last term to be equal to λQ, the following restriction must hold: $\alpha + \beta + \gamma + \delta = 1$. Hence, a test of constant returns to scale is equivalent to testing $H_0; \alpha + \beta + \gamma + \delta = 1$. The Cobb-Douglas production function is nonlinear in the variables, and can be linearized by taking logs of both sides, i.e.,

$$\log Q = \log A + \alpha \log K + \beta \log L + \gamma \log E + \delta \log M + u \tag{4.18}$$

This is a linear regression with $Y = \log Q$, $X_2 = \log K$, $X_3 = \log L$, $X_4 = \log E$ and $X_5 = \log M$. Ordinary least squares is BLUE on this non-linear model as long as u satisfies assumptions 1-4. Note that these disturbances entered the original Cobb-Douglas production function multiplicatively as $\exp(u_i)$. Had these disturbances entered additively as $Q = AK^{\alpha}L^{\beta}E^{\gamma}M^{\delta} + u$ then taking logs does not simplify the right hand side and one has to estimate this with nonlinear least squares, see Chapter 8. Now we can test constant returns to scale as follows. The unrestricted regression is given by (4.18) and its degrees of freedom are $n - 5$. Imposing H_0 means substituting the linear restriction by replacing say β by $(1 - \alpha - \gamma - \delta)$. This results after collecting terms in the following restricted regression with one less parameter

$$\log(Q/L) = \log A + \alpha \log(K/L) + \gamma \log(E/L) + \delta \log(M/L) + u \tag{4.19}$$

The degrees of freedom are $n - 4$. Once again all the ingredients for the test in (4.17) are there and this statistic is distributed as $F_1, n - 5$ under the null hypothesis.

Example 6: Joint significance of all the slope coefficients. The null hypothesis is

$$H_0; \beta_2 = \beta_3 = .. = \beta_K = 0$$

against the alternative H_1; at least one $\beta_k \neq 0$ for $k = 2, \ldots, K$. Under the null, only the constant is left in the regression. Problem 3.2 showed that for a regression of Y on a constant only, the least squares estimate of α is $\bar{Y}$. This means that the corresponding residual sum of squares is $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Therefore, $RRSS$ = Total sums of squares of regression (4.1) = $\Sigma_{i=1}^{n} y_i^2$.

The URSS is the usual residual sums of squares $\sum_{i=1}^{n} e_i^2$ from the unrestricted regression given by (4.1). Hence, the corresponding F-statistic for H_0 is

$$F = \frac{(TSS - RSS)/(K-1)}{RSS/(n-K)} = \frac{(\sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} e_i^2)/(K-1)}{\sum_{i=1}^{n} e_i^2/(n-K)} = \frac{R^2}{1-R^2} \cdot \frac{n-K}{K-1} \tag{4.20}$$

where $R^2 = 1 - (\sum_{i=1}^{n} e_i^2 / \sum_{i=1}^{n} y_i^2)$. This F-statistic has $(K-1)$ and $(n-K)$ degrees of freedom under H_0, and is usually reported by regression packages.

4.7 Dummy Variables

Many explanatory variables are qualitative in nature. For example, the head of a household could be male or female, white or non-white, employed or unemployed. In this case, one codes these variables as "M" for male and "F" for female, or change this qualitative variable into a quantitative variable called FEMALE which takes the value "0" for male and "1" for female. This obviously begs the question: "why not have a variable MALE that takes on the value 1 for male and 0 for female?" Actually, the variable MALE would be exactly 1-FEMALE. In other words, the zero and one can be thought of as a switch, which turns on when it is 1 and off when it is 0. Suppose that we are interested in the earnings of households, denoted by EARN, and MALE and FEMALE are the only explanatory variables available, then problem 10 asks the reader to verify that running OLS on the following model:

$$EARN = \alpha_M MALE + \alpha_F FEMALE + u \tag{4.21}$$

gives $\widehat{\alpha}_M$ = "average earnings of the males in the sample" and $\widehat{\alpha}_F$ = "average earnings of the females in the sample." Notice that there is no intercept in (4.21), this is because of what is known in the literature as the "dummy variable trap." Briefly stated, there will be perfect multicollinearity between MALE, FEMALE and the constant. In fact, MALE + FEMALE = 1. Some researchers may choose to include the intercept and exclude one of the sex dummy variables, say MALE, then

$$EARN = \alpha + \beta FEMALE + u \tag{4.22}$$

and the OLS estimates give $\widehat{\alpha}$ = "average earnings of males in the sample" = $\widehat{\alpha}_M$, while $\widehat{\beta} = \widehat{\alpha}_F - \widehat{\alpha}_M$ = "the difference in average earnings between females and males in the sample." Regression (4.22) is more popular when one is interested in contrasting the earnings between males and females and obtaining with one regression the markup or markdown in average earnings $(\widehat{\alpha}_F - \widehat{\alpha}_M)$ as well as the test of whether this difference is statistically different from zero. This would be simply the t-statistic on $\widehat{\beta}$ in (4.22). On the other hand, if one is interested in estimating the average earnings of males and females separately, then model (4.21) should be the one to consider. In this case, the t-test for $\widehat{\alpha}_F - \widehat{\alpha}_M = 0$ would involve further calculations not directly given from the regression in (4.21) but similar to the calculations given in Example 3.

What happens when another qualitative variable is included, to depict another classification of the individuals in the sample, say for example, race? If there are three race groups in the sample, WHITE, BLACK and HISPANIC. One could create a dummy variable for each of these classifications. For example, WHITE will take the value 1 when the individual is white

and 0 when the individual is non-white. Note that the dummy variable trap does not allow the inclusion of all three categories as they sum up to 1. Also, even if the intercept is dropped, once MALE and FEMALE are included, perfect multicollinearity is still present because MALE + FEMALE = WHITE + BLACK + HISPANIC. Therefore, one category from race should be dropped. Suits (1984) argues that the researcher should use the dummy variable category omission to his or her advantage, in interpreting the results, keeping in mind the purpose of the study. For example, if one is interested in comparing earnings across the sexes holding race constant, the omission of MALE or FEMALE is natural, whereas, if one is interested in the race differential in earnings holding gender constant, one of the race variables should be omitted. Whichever variable is omitted, this becomes the base category for which the other earnings are compared. Most researchers prefer to keep an intercept, although regression packages allow for a no intercept option. In this case one should omit one category from each of the race and sex classifications. For example, if MALE and WHITE are omitted:

$$EARN = \alpha + \beta_F FEMALE + \beta_B BLACK + \beta_H HISPANIC + u \tag{4.23}$$

Assuming the error u satisfies all the classical assumptions, and taking expected values of both sides of (4.23), one can see that the intercept α = the expected value of earnings of the omitted category which is "white males". For this category, all the other switches are off. Similarly, $\alpha + \beta_F$ is the expected value of earnings of "white females," since the FEMALE switch is on. One can conclude that β_F = difference in the expected value of earnings between white females and white males. Similarly, one can show that $\alpha + \beta_B$ is the expected earnings of "black males" and $\alpha + \beta_F + \beta_B$ is the expected earnings of "black females." Therefore, β_F represents the difference in expected earnings between black females and black males. In fact, problem 11 asks the reader to show that β_F represents the difference in expected earnings between hispanic females and hispanic males. In other words, β_F represents the differential in expected earnings between females and males holding race constant. Similarly, one can show that β_B is the difference in expected earnings between blacks and whites holding sex constant, and β_H is the differential in expected earnings between hispanics and whites holding sex constant. The main key to the interpretation of the dummy variable coefficients is to be able to turn on and turn off the proper switches, and write the correct expectations.

The real regression will contain other quantitative and qualitative variables, like

$$\begin{aligned} EARN = {} & \alpha + \beta_F FEMALE + \beta_B BLACK + \beta_H HISPANIC + \gamma_1 EXP \\ & + \gamma_2 EXP^2 + \gamma_3 EDUC + \gamma_4 UNION + u \end{aligned} \tag{4.24}$$

where EXP is years of job experience, EDUC is years of education, and UNION is 1 if the individual belongs to a union and 0 otherwise. EXP^2 is the squared value of EXP. Once again, one can interpret the coefficients of these regressions by turning on or off the proper switches. For example, γ_4 is interpreted as the expected difference in earnings between union and non-union members holding all other variables included in (4.24) constant. Halvorsen and Palmquist (1980) warn economists about the interpretation of dummy variable coefficients when the dependent variable is in logs. For example, if the earnings equation is semi-logarithmic:

$$\log(Earnings) = \alpha + \beta UNION + \gamma EDUC + u$$

then γ = % change in earnings for one extra year of education, holding union membership constant. But, what about the returns for union membership? If we let Y_1 = log(Earnings)

when the individual belongs to a union, and $Y_0 = \log(\text{Earnings})$ when the individual does not belong to a union, then $g = \%$ change in earnings due to union membership $= (e^{Y_1} - e^{Y_0})/e^{Y_0}$. Equivalently, one can write that $\log(1+g) = Y_1 - Y_0 = \beta$, or that $g = e^{\beta} - 1$. In other words, one should not hasten to conclude that β has the same interpretation as γ. In fact, the % change in earnings due to union membership is $e^{\beta} - 1$ and not β. The error involved in using $\widehat{\beta}$ rather than $e^{\widehat{\beta}} - 1$ to estimate g could be substantial, especially if $\widehat{\beta}$ is large. For example, when $\widehat{\beta} = 0.5, 0.75, 1$; $\widehat{g} = e^{\widehat{\beta}} - 1 = 0.65, 1.12, 1.72$, respectively. Kennedy (1981) notes that if $\widehat{\beta}$ is unbiased for β, $\widehat{g}$ is not necessarily unbiased for g. However, consistency of $\widehat{\beta}$ implies consistency for $\widehat{g}$. If one assumes log-normal distributed errors, then $E(e^{\widehat{\beta}}) = e^{\beta + 0.5\text{Var}(\widehat{\beta})}$. Based on this result, Kennedy (1981) suggests estimating g by $\widetilde{g} = e^{\widehat{\beta} + 0.5\widehat{\text{Var}}(\widehat{\beta})}$-1, where $\widehat{\text{Var}}(\widehat{\beta})$ is a consistent estimate of $\text{Var}(\widehat{\beta})$.

Another use of dummy variables is in taking into account seasonal factors, i.e., including 3 seasonal dummy variables with the omitted season becoming the base for comparison.[1] For example:

$$Sales = \alpha + \beta_W Winter + \beta_S Spring + \beta_F Fall + \gamma_1 Price + u \tag{4.25}$$

the omitted season being the Summer season, and if (4.25) models the sales of air-conditioning units, then β_F is the difference in expected sales between the Fall and Summer seasons, holding the price of an air-conditioning unit constant. If these were heating units one may want to change the base season for comparison.

Another use of dummy variables is for War years, where consumption is not at its normal level say due to rationing. Consider estimating the following consumption function

$$C_t = \alpha + \beta Y_t + \delta \text{WAR}_t + u_t \quad t = 1, 2, \ldots, T \tag{4.26}$$

where C_t denotes real per capita consumption, Y_t denotes real per capita personal disposable income, and WAR_t is a dummy variable taking the value 1 if it is a War time period and 0 otherwise. Note that the War years do not affect the slope of the consumption line with respect to income, only the intercept. The intercept is α in non-War years and $\alpha + \delta$ in War years. In other words, the marginal propensity out of income is the same in War and non-War years, only the level of consumption is different.

Of course, one can dummy other unusual years like periods of strike, years of natural disaster, earthquakes, floods, hurricanes, or external shocks beyond control, like the oil embargo of 1973. If this dummy includes only one year like 1973, then the dummy variable for 1973, call it D_{73}, takes the value 1 for 1973 and zero otherwise. Including D_{73} as an extra variable in the regression has the effect of removing the 1973 observation from estimation purposes, and the resulting regression coefficients estimates are exactly the same as those obtained excluding the 1973 observation and its corresponding dummy variable. In fact, using matrix algebra in Chapter 7, we will show that the coefficient estimate of D_{73} is the forecast error for 1973, using the regression that ignores the 1973 observations. In addition, the standard error of the dummy coefficient estimates is the standard error of this forecast. This is a much easier way of obtaining the forecast error and its standard error from the regression package without additional computations, see Salkever (1976). More on this in Chapter 7.

Interaction Effects

So far the dummy variables have been used to shift the intercept of the regression keeping the slopes constant. One can also use the dummy variables to shift the slopes by letting them interact with the explanatory variables. For example, consider the following earnings equation:

$$EARN = \alpha + \alpha_F FEMALE + \beta EDUC + u \tag{4.27}$$

In this regression, only the intercept shifts from males to females. The returns to an extra year of education is simply β, which is assumed to be the same for males as well as females. But if we now introduce the interaction variable (FEMALE $\times$ EDUC), then the regression becomes:

$$EARN = \alpha + \alpha_F FEMALE + \beta EDUC + \gamma(FEMALE \times EDUC) + u \tag{4.28}$$

In this case, the returns to an extra year of education depends upon the sex of the individual. In fact, $\partial(EARN)/\partial(EDUC) = \beta + \gamma(FEMALE) = \beta$ if male, and $\beta + \gamma$ if female. Note that the interaction variable = EDUC if the individual is female and 0 if the individual is male.

Estimating (4.28) is equivalent to estimating two earnings equations, one for males and another one for females, separately. The only difference is that (4.28) imposes the same variance across the two groups, whereas separate regressions do not impose this, albeit restrictive, equality of the variances assumption. This set-up is ideal for testing the equality of slopes, equality of intercepts, or equality of both intercepts and slopes across the sexes. This can be done with the F-test described in (4.17). In fact, for H_0; equality of slopes, given different intercepts, the restricted residuals sum of squares (RRSS) is obtained from (4.27), while the unrestricted residuals sum of squares (URSS) is obtained from (4.28). Problem 12 asks the reader to set up the F-test for the following null hypothesis: (i) equality of slopes and intercepts, and (ii) equality of intercepts given the same slopes.

Dummy variables have many useful applications in economics. For example, several tests including the Chow (1960) test, and Utts (1982) Rainbow test described in Chapter 8, can be applied using dummy variable regressions. Additionally, they can be used in modeling splines, see Poirier (1976) and Suits, Mason and Chan (1978), and fixed effects in panel data, see Chapter 12. Finally, when the dependent variable is itself a dummy variable, the regression equation needs special treatment, see Chapter 13 on qualitative limited dependent variables.

Empirical Example: Table 4.1 gives the results of a regression on 595 individuals drawn from the Panel Study of Income Dynamics (PSID) in 1982. This data is provided on the Springer web site as EARN.ASC. A description of the data is given in Cornwell and Rupert (1988). In particular, log wage is regressed on years of education (ED), weeks worked (WKS), years of full-time work experience (EXP), occupation (OCC = 1, if the individual is in a blue-collar occupation), residence (SOUTH = 1, SMSA = 1, if the individual resides in the South, or in a standard metropolitan statistical area), industry (IND = 1, if the individual works in a manufacturing industry), marital status (MS = 1, if the individual is married), sex and race (FEM = 1, BLK = 1, if the individual is female or black), union coverage (UNION = 1, if the individual's wage is set by a union contract). These results show that the returns to an extra year of schooling is 5.7%, holding everything else constant. It shows that Males on the average earn more than Females. Blacks on the average earn less than Whites, and Union workers earn more than non-union workers. Individuals residing in the South earn less than those living elsewhere. Those residing in a standard metropolitan statistical area earn more on the average than those

Table 4.1 Earnings Regression for 1982

Dependent Variable: LWAGE
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob > F
Model	12	52.48064	4.37339	41.263	0.0001
Error	582	61.68465	0.10599		
C Total	594	114.16529			

Root MSE	0.32556	R-square	0.4597
Dep Mean	6.95074	Adj R-sq	0.4485
C.V.	4.68377		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	5.590093	0.19011263	29.404	0.0001
WKS	1	0.003413	0.00267762	1.275	0.2030
SOUTH	1	–0.058763	0.03090689	–1.901	0.0578
SMSA	1	0.166191	0.02955099	5.624	0.0001
MS	1	0.095237	0.04892770	1.946	0.0521
EXP	1	0.029380	0.00652410	4.503	0.0001
EXP2	1	–0.000486	0.00012680	–3.833	0.0001
OCC	1	–0.161522	0.03690729	–4.376	0.0001
IND	1	0.084663	0.02916370	2.903	0.0038
UNION	1	0.106278	0.03167547	3.355	0.0008
FEM	1	–0.324557	0.06072947	–5.344	0.0001
BLK	1	–0.190422	0.05441180	–3.500	0.0005
ED	1	0.057194	0.00659101	8.678	0.0001

who do not. Individuals who work in a manufacturing industry or are not blue collar workers or are married earn more on the average than those who are not. For $EXP2 = (EXP)^2$, this regression indicates a significant quadratic relationship between earnings and experience. All the variables were significant at the 5% level except for WKS, SOUTH and MS.

Note

1. There are more sophisticated ways of seasonal adjustment than introducing seasonal dummies, see Judge et al. (1985).

Problems

1. For the Cigarette Data given in Table 3.2. Run the following regressions:
   (a) Real per capita consumption of cigarettes on real price and real per capita income. (All variables are in log form, and all regressions in this problem include a constant).

(b) Real per capita consumption of cigarettes on real price.

(c) Real per capita income on real price.

(d) Real per capita consumption on the residuals of part (c).

(e) Residuals from part (b) on the residuals in part (c).

(f) Compare the regression slope estimates in parts (d) and (e) with the regression coefficient estimate of the real income coefficient in part (a), what do you conclude?

2. *Simple versus Multiple Regression Coefficients.* This is based on Baltagi (1987b). Consider the multiple regression

$$Y_i = \alpha + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad i = 1, 2, \ldots, n$$

along with the following auxiliary regressions:

$$\begin{aligned} X_{2i} &= \widehat{a} + \widehat{b} X_{3i} + \widehat{\nu}_{2i} \\ X_{3i} &= \widehat{c} + \widehat{d} X_{2i} + \widehat{\nu}_{3i} \end{aligned}$$

In section 4.3, we showed that $\widehat{\beta}_2$, the OLS estimate of β_2 can be interpreted as a simple regression of Y on the OLS residuals $\widehat{\nu}_2$. A similar interpretation can be given to $\widehat{\beta}_3$. Kennedy (1981, p. 416) claims that $\widehat{\beta}_2$ is not necessarily the same as $\widehat{\delta}_2$, the OLS estimate of δ_2 obtained from the regression Y on $\widehat{\nu}_2$, $\widehat{\nu}_3$ and a constant, $Y_i = \gamma + \delta_2 \widehat{\nu}_{2i} + \delta_3 \widehat{\nu}_{3i} + w_i$. Prove this claim by finding a relationship between the $\widehat{\beta}$'s and the $\widehat{\delta}$'s.

3. For the simple regression $Y_i = \alpha + \beta X_i + u_i$ considered in Chapter 3, show that

(a) $\widehat{\beta}_{OLS} = \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2$ can be obtained using the residual interpretation by regressing X on a constant first, getting the residuals $\widehat{\nu}$ and then regressing Y on $\widehat{\nu}$.

(b) $\widehat{\alpha}_{OLS} = \bar{Y} - \widehat{\beta}_{OLS} \bar{X}$ can be obtained using the residual interpretation by regressing 1 on X and obtaining the residuals $\widehat{\omega}$ and then regressing Y on $\widehat{\omega}$.

(c) Check the var$(\widehat{\alpha}_{OLS})$ and var$(\widehat{\beta}_{OLS})$ in parts (a) and (b) with those obtained from the residualing interpretation.

4. *Effect of Additional Regressors on* R^2. This is based on Nieswiadomy (1986).

(a) Suppose that the multiple regression given in (4.1) has K_1 regressors in it. Denote the least squares sum of squared errors by SSE_1. Now add K_2 regressors so that the total number of regressors is $K = K_1 + K_2$. Denote the corresponding least squares sum of squared errors by SSE_2. Show that $SSE_2 \leq SSE_1$, and conclude that the corresponding R-squares satisfy $R_2^2 \geq R_1^2$.

(b) Derive the equality given in (4.16) starting from the definition of R^2 and $\bar{R}^2$.

(c) Show that the corresponding $\bar{R}$-squares satisfy $\bar{R}_1^2 \geq \bar{R}_2^2$ when the F-statistic for the joint significance of these additional K_2 regressors is less than or equal to one.

5. Let Y be the output and X_2 = skilled labor and X_3 = unskilled labor in the following relationship:

$$Y_i = \alpha + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 (X_{2i} + X_{3i}) + \beta_5 X_{2i}^2 + \beta_6 X_{3i}^2 + u_i$$

What parameters are estimable by OLS?

6. Suppose that we have estimated the parameters of the multiple regression model:

$$Y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + u_t$$

by *Ordinary Least Squares* (OLS) method. Denote the estimated residuals by $(e_t, t = 1, \ldots, T)$ and the predicted values by $(\widehat{Y}_t, t = 1, \ldots, T)$.

(a) What is the R^2 of the regression of e on a constant, X_2 and X_3?

(b) If we regress Y on a constant and $\widehat{Y}$, what are the estimated intercept and slope coefficients? What is the relationship between the R^2 of this regression and the R^2 of the original regression?

(c) If we regress Y on a constant and e, what are the estimated intercept and slope coefficients? What is the relationship between the R^2 of this regression and the R^2 of the original regression?

(d) Suppose that we add a new explanatory variable X_4 to the original model and re-estimate the parameters by OLS. Show that the estimated coefficient of X_4 and its estimated standard error will be the same as in the OLS regression of e on a constant, X_2, X_3 and X_4.

7. Consider the Cobb-Douglas production function in example 5. How can you test for constant returns to scale using a t-statistic from the unrestricted regression given in (4.18).

8. For the multiple regression given in (4.1). Set up the F-statistic described in (4.17) for testing

(a) $H_0; \beta_2 = \beta_4 = \beta_6$.

(b) $H_0; \beta_2 = -\beta_3$ and $\beta_5 - \beta_6 = 1$.

9. *Monte Carlo Experiments.* Hanushek and Jackson (1977, pp. 60-65) generated the following data $Y_i = 15 + 1X_{2i} + 2X_{3i} + u_i$ for $i = 1, 2, \ldots, 25$ with a fixed set of X_{2i} and X_{3i}, and u_i's that are IID $\sim N(0, 100)$. For each set of 25 u_i's drawn randomly from the normal distribution, a corresponding set of 25 Y_i's are created from the above equation. Then OLS is performed on the resulting data set. This can be repeated as many times as we can afford. 400 replications were performed by Hanushek and Jackson. This means that they generated 400 data sets each of size 25 and ran 400 regressions giving 400 OLS estimates of α, β_2, β_3 and σ^2. The classical assumptions are satisfied for this model, by construction, so we expect these OLS estimators to be BLUE, MLE and efficient.

(a) Replicate the Monte Carlo experiments of Hanushek and Jackson (1977) and generate the means of the 400 estimates of the regression coefficients as well as σ^2. Are these estimates unbiased?

(b) Compute the standard deviation of these 400 estimates and call this $\widehat{\sigma}_b$. Also compute the average of the 400 standard errors of the regression estimates reported by the regression. Denote this mean by $\overline{s}_b$. Compare these two estimates of the standard deviation of the regression coefficient estimates to the true standard deviation knowing the true σ^2. What do you conclude?

(c) Plot the frequency of these regression coefficients estimates? Does it resemble its theoretical distribution.

(d) Increase the sample size form 25 to 50 and repeat the experiment. What do you observe?

10. (a) Derive the OLS estimates of α_F and α_M for $Y_i = \alpha_F F_i + \alpha_M M_i + u_i$ where Y is Earnings, F is FEMALE and M is MALE, see (4.21). Show that $\widehat{\alpha}_F = \bar{Y}_F$, the average of the Y_i's only for females, and $\widehat{\alpha}_M = \bar{Y}_M$, the average of the Y_i's only for males.

(b) Suppose that the regression is $Y_i = \alpha + \beta F_i + u_i$, see (4.22). Show that $\widehat{\alpha} = \widehat{\alpha}_M$, and $\widehat{\beta} = \widehat{\alpha}_F - \widehat{\alpha}_M$.

(c) Substitute $M = 1 - F$ in (4.21) and show that $\alpha = \alpha_M$ and $\beta = \alpha_F - \alpha_M$.

(d) Verify parts (a), (b) and (c) using the earnings data underlying Table 4.1.

11. For equation (4.23)

$$EARN = \alpha + \beta_F FEMALE + \beta_B BLACK + \beta_H HISPANIC + u$$

Show that

(a) E(Earnings/Hispanic Female) $= \alpha + \beta_F + \beta_H$; also E(Earnings/Hispanic Male) $= \alpha + \beta_H$. Conclude that β_F = E(Earnings/Hispanic Female) – E(Earnings/Hispanic Male).

(b) E(Earnings/Hispanic Female) – E(Earnings/White Female) = E(Earnings/Hispanic Male) – E(Earnings/White Male) $= \beta_H$.

(c) E(Earnings/Black Female) – E(Earnings/White Female) = E(Earnings/Black Male) – E(Earnings/White Male) $= \beta_B$.

12. For the earnings equation given in (4.28), how would you set up the F-test and what are the restricted and unrestricted regressions for testing the following hypotheses:

(a) The equality of slopes and intercepts for Males and Females.

(b) The equality of intercepts given the same slopes for Males and Females. Show that the resulting F-statistic is the square of a t-statistic from the unrestricted regression.

(c) The equality of intercepts allowing for different slopes for Males and Females. Show that the resulting F-statistic is the square of a t-statistic from the unrestricted regression.

(d) Apply your results in parts (a), (b) and (c) to the earnings data underlying Table 4.1.

13. For the earnings data regression underlying Table 4.1.

(a) Replicate the regression results given in Table 4.1.

(b) Verify that the joint significance of all slope coefficients can be obtained from (4.20).

(c) How would you test the joint restriction that expected earnings are the same for Males and Females whether Black or Non-Black holding everything else constant?

(d) How would you test the joint restriction that expected earnings are the same whether the individual is married or not and whether this individual belongs to a Union or not?

(e) From Table 4.1 what is your estimate of the % change in earnings due to Union membership? If the disturbances are assumed to be log-normal, what would be the estimate suggested by Kennedy (1981) for this % change in earnings?

(f) What is your estimate of the % change in earnings due to the individual being married?

14. *Crude Quality.* Using the data set of U.S. oil field postings on crude prices ($/barrel), gravity (degree API) and sulphur (% sulphur) given in the CRUDES.ASC file on the Springer web site.

(a) Estimate the following multiple regression model: POIL $= \beta_1 + \beta_2$GRAVITY $+ \beta_3$ SULPHUR $+ \epsilon$.

(b) Regress GRAVITY $= \alpha_0 + \alpha_1$SULPHUR $+ \nu_t$ then compute the residuals ($\widehat{\nu}_t$). Now perform the regression

$$POIL = \gamma_1 + \gamma_2 \widehat{\nu}_t + \epsilon$$

Verify that $\widehat{\gamma}_2$ is the same as $\widehat{\beta}_2$ in part (a). What does this tell you?

(c) Regress POIL $= \phi_1 + \phi_2$SULPHUR $+ w$. Compute the residuals ($\widehat{w}$). Now regress $\widehat{w}$ on $\widehat{\nu}$ obtained from part (b), to get $\widehat{w}_t = \widehat{\delta}_1 + \widehat{\delta}_2 \widehat{\nu}_t +$ residuals. Show that $\widehat{\delta}_2 = \widehat{\beta}_2$ in part (a). Again, what does this tell you?

(d) To illustrate how additional data affects multicollinearity, show how your regression in part (a) changes when the sample is restricted to the first 25 crudes.

(e) Delete all crudes with sulphur content outside the range of 1 to 2 percent and run the multiple regression in part (a). Discuss and interpret these results.

15. Consider the U.S. gasoline data from 1950-1987 given in Table 4.2, and obtained from the file USGAS.ASC on the Springer web site.

Table 4.2 U.S. Gasoline Data: 1950–1987

Year	CAR	QMG (1,000 Gallons)	PMG ($)	POP (1,000)	RGNP (Billion)	PGNP
1950	49195212	40617285	0.272	152271	1090.4	26.1
1951	51948796	43896887	0.276	154878	1179.2	27.9
1952	53301329	46428148	0.287	157553	1226.1	28.3
1953	56313281	49374047	0.290	160184	1282.1	28.5
1954	58622547	51107135	0.291	163026	1252.1	29.0
1955	62688792	54333255	0.299	165931	1356.7	29.3
1956	65153810	56022406	0.310	168903	1383.5	30.3
1957	67124904	57415622	0.304	171984	1410.2	31.4
1958	68296594	59154330	0.305	174882	1384.7	32.1
1959	71354420	61596548	0.311	177830	1481.0	32.6
1960	73868682	62811854	0.308	180671	1517.2	33.2
1961	75958215	63978489	0.306	183691	1547.9	33.6
1962	79173329	62531373	0.304	186538	1647.9	34.0
1963	82713717	64779104	0.304	189242	1711.6	34.5
1964	86301207	67663848	0.312	191889	1806.9	35.0
1965	90360721	70337126	0.321	194303	1918.5	35.7
1966	93962030	73638812	0.332	196560	2048.9	36.6
1967	96930949	76139326	0.337	198712	2100.3	37.8
1968	101039113	80772657	0.348	200706	2195.4	39.4
1969	103562018	85416084	0.357	202677	2260.7	41.2
1970	106807629	88684050	0.364	205052	2250.7	43.4
1971	111297459	92194620	0.361	207661	2332.0	45.6
1972	117051638	95348904	0.388	209896	2465.5	47.5
1973	123811741	99804600	0.524	211909	2602.8	50.2
1974	127951254	100212210	0.572	213854	2564.2	55.1
1975	130918918	102327750	0.595	215973	2530.9	60.4
1976	136333934	106972740	0.631	218035	2680.5	63.5
1977	141523197	110023410	0.657	220239	2822.4	67.3
1978	146484336	113625960	0.678	222585	3115.2	72.2
1979	149422205	107831220	0.857	225055	3192.4	78.6
1980	153357876	100856070	1.191	227757	3187.1	85.7
1981	155907473	100994040	1.311	230138	3248.8	94.0
1982	156993694	100242870	1.222	232520	3166.0	100.0
1983	161017926	101515260	1.157	234799	3279.1	103.9
1984	163432944	102603690	1.129	237001	3489.9	107.9
1985	168743817	104719230	1.115	239279	3585.2	111.5
1986	173255850	107831220	0.857	241613	3676.5	114.5
1987	177922000	110467980	0.897	243915	3847.0	117.7

CAR: Stock of Cars
RMG: Motor Gasoline Consumption
PMG: Retail Price of Motor Gasoline
POP: Population
RGNP: Real GNP in 1982 dollars
PGNP: GNP Deflator (1982=100)

(a) For the period 1950-1972 estimate models (1) and (2):

$$\begin{aligned} \log QMG &= \beta_1 + \beta_2 \log CAR + \beta_3 \log POP + \beta_4 \log RGNP \\ &\quad + \beta_5 \log PGNP + \beta_6 \log PMG + u \end{aligned} \tag{1}$$

$$\log \frac{QMG}{CAR} = \gamma_1 + \gamma_2 \log \frac{RGNP}{POP} + \gamma_3 \log \frac{CAR}{POP} + \gamma_4 \log \frac{PMG}{PGNP} + \nu \tag{2}$$

(b) What restrictions should the β's satisfy in model (1) in order to yield the γ's in model (2)?

(c) Compare the estimates and the corresponding standard errors from models (1) and (2).

(d) Compute the simple correlations among the X's in model (1). What do you observe?

(e) Use the Chow-F test to test the parametric restrictions obtained in part (b).

(f) Estimate equations (1) and (2) now using the full data set 1950-1987. Discuss briefly the effects on individual parameter estimates and their standard errors of the larger data set.

(g) Using a dummy variable, test the hypothesis that gasoline demand per CAR permanently shifted downward for model (2) following the Arab Oil Embargo in 1973?

(h) Construct a dummy variable regression that will test whether the price elasticity has changed after 1973.

16. Consider the following model for the demand for natural gas by residential sector, call it model (1):

$$\log Cons_{it} = \beta_0 + \beta_1 \log Pg_{it} + \beta_2 \log Po_{it} + \beta_3 \log Pe_{it} + \beta_4 \log HDD_{it} + \beta_5 \log PI_{it} + u_{it}$$

where $i = 1, 2, \ldots, 6$ states and $t = 1, 2, \ldots, 23$ years. Cons is the consumption of natural gas by residential sector, Pg, Po and Pe are the prices of natural gas, distillate fuel oil, and electricity of the residential sector. HDD is heating degree days and PI is real per capita personal income. The data covers 6 states: NY, FL, MI, TX, UT and CA over the period 1967-1989. It is given in the NATURAL.ASC file on the Springer web site.

(a) Estimate the above model by OLS. Call this model (1). What do the parameter estimates imply about the relationship between the fuels?

(b) Plot actual consumption versus the predicted values. What do you observe?

(c) Add a dummy variable for each state except California and run OLS. Call this model (2). Compute the parameter estimates and standard errors and compare to model (1). Do any of the interpretations of the price coefficients change? What is the interpretation of the New York dummy variable? What is the predicted consumption of natural gas for New York in 1989?

(d) Test the hypothesis that the intercepts of New York and California are the same.

(e) Test the hypothesis that **all** the states have the same intercept.

(f) Add a dummy variable for each state and run OLS without an intercept. Call this model (3). Compare the parameter estimates and standard errors to the first two models. What is the interpretation of the coefficient of the New York dummy variable? What is the predicted consumption of natural gas for New York in 1989?

(g) Using the regression in part (f), test the hypothesis that the intercepts of New York and California are the same.

References

This chapter draws upon the material in Kelejian and Oates (1989) and Wallace and Silver (1988). Several econometrics books have an excellent discussion on dummy variables, see Gujarati (1978), Judge et al. (1985), Kennedy (1992), Johnston (1984) and Maddala (2001), to mention a few. Other readings referenced in this chapter include:

Baltagi, B.H. (1987a), "To Pool or Not to Pool: The Quality Bank Case," *The American Statistician*, 41: 150-152.

Baltagi, B.H. (1987b), "Simple versus Multiple Regression Coefficients," *Econometric Theory*, Problem 87.1.1, 3: 159.

Chow, G.C. (1960), "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, 28: 591-605.

Cornwell, C. and P. Rupert (1988), "Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variables Estimators," *Journal of Applied Econometrics*, 3: 149-155.

Dufour, J.M. (1980), "Dummy Variables and Predictive Tests for Structural Change," *Economics Letters*, 6: 241-247.

Dufour, J.M. (1982), "Recursive Stability of Linear Regression Relationships," *Journal of Econometrics*, 19: 31-76.

Gujarati, D. (1970), "Use of Dummy Variables in Testing for Equality Between Sets of Coefficients in Two Linear Regressions: A Note," *The American Statistician*, 24: 18-21.

Gujarati, D. (1970), "Use of Dummy Variables in Testing for Equality Between Sets of Coefficients in Two Linear Regressions: A Generalization," *The American Statistician*, 24: 50-52.

Halvorsen, R. and R. Palmquist (1980), "The Interpretation of Dummy Variables in Semilogarithmic Equations," *American Economic Review*, 70: 474-475.

Hanushek, E.A. and J.E. Jackson (1977), *Statistical Methods for Social Scientists* (Academic Press: New York).

Hill, R. Carter and L.C. Adkins (2001), "Collinearity," Chapter 12 in B.H. Baltagi (ed.) *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Kennedy, P.E. (1981), "Estimation with Correctly Interpreted Dummy Variables in Semilogarithmic Equations," *American Economic Review*, 71: 802.

Kennedy, P.E. (1981), "The Balentine: A Graphical Aid for Econometrics," *Australian Economic Papers*, 20: 414-416.

Kennedy, P.E. (1986), "Interpreting Dummy Variables," *Review of Economics and Statistics*, 68: 174-175.

Nieswiadomy, M. (1986), "Effect of an Additional Regressor on R^2," *Econometric Theory*, Problem 86.3.1, 2:442.

Poirier, D. (1976), *The Econometrics of Structural Change*(North Holland: Amsterdam).

Salkever, D. (1976), "The Use of Dummy Variables to Compute Predictions, Prediction Errors, and Confidence Intervals," *Journal of Econometrics*, 4: 393-397.

Suits, D. (1984), "Dummy Variables: Mechanics vs Interpretation," *Review of Economics and Statistics*, 66: 132-139.

Suits, D.B., A. Mason and L. Chan (1978), "Spline Functions Fitted by Standard Regression Methods," *Review of Economics and Statistics*, 60: 132-139.

Utts, J. (1982), "The Rainbow Test for Lack of Fit in Regression," *Communications in Statistics-Theory and Methods*, 11: 1801-1815.

Appendix
Residual Interpretation of Multiple Regression Estimates

Proof of Claim 1: Regressing X_2 on all the other X's yields residuals $\widehat{\nu}_2$ that satisfy the usual properties of OLS residuals similar to those in (4.2), i.e.,

$$\textstyle\sum_{i=1}^n \widehat{\nu}_{2i} = 0, \;\; \sum_{i=1}^n \widehat{\nu}_{2i} X_{3i} = \sum_{i=1}^n \widehat{\nu}_{2i} X_{4i} = .. = \sum_{i=1}^n \widehat{\nu}_{2i} X_{Ki} = 0 \tag{A.1}$$

Note that X_2 is the dependent variable of this regression, and $\widehat{X}_2$ is the predicted value from this regression. The latter satisfies $\sum_{i=1}^n \widehat{\nu}_{2i}\widehat{X}_{2i} = 0$. This holds because $\widehat{X}_2$ is a linear combination of the other X's, all of which satisfy (A.1). Turn now to the estimated regression equation:

$$Y_i = \widehat{\alpha} + \widehat{\beta}_2 X_{2i} + .. + \widehat{\beta}_K X_{Ki} + e_i \tag{A.2}$$

Multiply (A.2) by X_{2i} and sum

$$\textstyle\sum_{i=1}^n X_{2i} Y_i = \widehat{\alpha} \sum_{i=1}^n X_{2i} + \widehat{\beta}_2 \sum_{i=1}^n X_{2i}^2 + .. + \widehat{\beta}_K \sum_{i=1}^n X_{2i} X_{Ki} \tag{A.3}$$

This uses the fact that $\sum_{i=1}^n X_{2i} e_i = 0$. Alternatively, (A.3) is just the second equation from (4.3). Substituting $X_{2i} = \widehat{X}_{2i} + \widehat{\nu}_{2i}$, in (A.3) one gets

$$\begin{aligned}\textstyle\sum_{i=1}^n \widehat{X}_{2i} Y_i + \sum_{i=1}^n \widehat{\nu}_{2i} Y_i = \widehat{\alpha} \sum_{i=1}^n \widehat{X}_{2i} + \widehat{\beta}_2 \sum_{i=1}^n \widehat{X}_{2i}^2 + .. \\ + \;\; \widehat{\beta}_K \textstyle\sum_{i=1}^n \widehat{X}_{2i} X_{Ki} + \widehat{\beta}_2 \sum_{i=1}^n \widehat{\nu}_{2i}^2\end{aligned} \tag{A.4}$$

using (A.1) and the fact that $\Sigma_{i=1}^n \widehat{X}_{2i}\widehat{\nu}_{2i} = 0$. Multiply (A.2) by $\widehat{X}_{2i}$ and sum, we get

$$\textstyle\sum_{i=1}^n \widehat{X}_{2i} Y_i = \widehat{\alpha} \sum_{i=1}^n \widehat{X}_{2i} + \widehat{\beta}_2 \sum_{i=1}^n \widehat{X}_{2i} X_{2i} + .. + \widehat{\beta}_K \sum_{i=1}^n \widehat{X}_{2i} X_{Ki} + \sum_{i=1}^n \widehat{X}_{2i} e_i \tag{A.5}$$

But $\sum_{i=1}^n \widehat{X}_2 e_i = 0$ since $\widehat{X}_2$ is a linear combination of all the other X's, all of which satisfy (4.2). Also, $\sum_{i=1}^n \widehat{X}_{2i} X_{2i} = \sum_{i=1}^n \widehat{X}_{2i}^2$ since $\sum_{i=1}^n \widehat{X}_{2i}\widehat{\nu}_{2i} = 0$. Hence (A.5) reduces to

$$\textstyle\sum_{i=1}^n \widehat{X}_{2i} Y_i = \widehat{\alpha} \sum_{i=1}^n \widehat{X}_{2i} + \widehat{\beta}_2 \sum_{i=1}^n \widehat{X}_{2i}^2 + .. + \widehat{\beta}_K \sum_{i=1}^n \widehat{X}_{2i} X_{Ki} \tag{A.6}$$

Subtracting (A.6) from (A.4), we get

$$\textstyle\sum_{i=1}^n \widehat{\nu}_{2i} Y_i = \widehat{\beta}_2 \sum_{i=1}^n \widehat{\nu}_{2i}^2 \tag{A.7}$$

and $\widehat{\beta}_2$ is the slope estimate of the simple regression of Y on $\widehat{\nu}_2$ as given in (4.5).

By substituting for Y_i its expression from equation (4.1) in (4.5) we get

$$\widehat{\beta}_2 = \beta_2 \textstyle\sum_{i=1}^n X_{2i}\widehat{\nu}_{2i} / \sum_{i=1}^n \widehat{\nu}_{2i}^2 + \sum_{i=1}^n \widehat{\nu}_{2i} u_i / \sum_{i=1}^n \widehat{\nu}_{2i}^2 \tag{A.8}$$

where $\sum_{i=1}^n X_{1i}\widehat{\nu}_{2i} = 0$ and $\sum_{i=1}^n \widehat{\nu}_{2i} = 0$. But, $X_{2i} = \widehat{X}_{2i} + \widehat{\nu}_{2i}$ and $\sum_{i=1}^n \widehat{X}_{2i}\widehat{\nu}_{2i} = 0$, which implies that $\sum_{i=1}^n X_{2i}\widehat{\nu}_{2i} = \sum_{i=1}^n \widehat{\nu}_{2i}^2$ and $\widehat{\beta}_2 = \beta_2 + \sum_{i=1}^n \widehat{\nu}_{2i} u_i / \sum_{i=1}^n \widehat{\nu}_{2i}^2$. This means that $\widehat{\beta}_2$ is unbiased with

$E(\widehat{\beta}_2) = \beta_2$ since $\widehat{\nu}_2$ is a linear combination of the X's and these in turn are not correlated with the u's. Also,

$$\text{var}(\widehat{\beta}_2) = E(\widehat{\beta}_2 - \beta_2)^2 = E(\textstyle\sum_{i=1}^n \widehat{\nu}_{2i}u_i / \sum_{i=1}^n \widehat{\nu}_{2i}^2)^2 = \sigma^2 / \sum_{i=1}^n \widehat{\nu}_{2i}^2$$

The same results apply for any $\widehat{\beta}_k$ for $k = 2, \ldots, K$, i.e.,

$$\widehat{\beta}_k = \textstyle\sum_{i=1}^n \widehat{\nu}_{ki}Y_i / \sum_{i=1}^n \widehat{\nu}_{ki}^2 \tag{A.9}$$

where $\widehat{\nu}_k$ is the OLS residual of X_k on all the other X's in the regression. Similarly,

$$\widehat{\beta}_k = \beta_k + \textstyle\sum_{i=1}^n \widehat{\nu}_{ki}u_i / \sum_{i=1}^n \widehat{\nu}_{ki}^2 \tag{A.10}$$

and $E(\widehat{\beta}_k) = \beta_k$ with $\text{var}(\widehat{\beta}_k) = \sigma^2 / \sum_{i=1}^n \widehat{\nu}_{ki}^2$ for $k = 2, \ldots, K$. Note also that

$$\begin{aligned} \text{cov}(\widehat{\beta}_2, \widehat{\beta}_k) &= E(\widehat{\beta}_2 - \beta_2)(\widehat{\beta}_k - \beta_k) = E(\textstyle\sum_{i=1}^n \widehat{\nu}_{2i}u_i / \sum_{i=1}^n \widehat{\nu}_{2i}^2)(\sum_{i=1}^n \widehat{\nu}_{ki}u_i / \sum_{i=1}^n \widehat{\nu}_{ki}^2) \\ &= \sigma^2 \textstyle\sum_{i=1}^n \widehat{\nu}_{2i}\widehat{\nu}_{ki} / \sum_{i=1}^n \widehat{\nu}_{2i}^2 \sum_{i=1}^n \widehat{\nu}_{ki}^2 \end{aligned}$$

Proof of Claim 2: Regressing Y on all the other X's yields, $Y_i = \widetilde{Y}_i + \widetilde{\omega}_i$. Substituting this expression for Y_i in (4.5) one gets

$$\widehat{\beta}_2 = (\textstyle\sum_{i=1}^n \widehat{\nu}_{2i}\widetilde{Y}_i + \sum_{i=1}^n \widehat{\nu}_{2i}\widetilde{\omega}_i) / \sum_{i=1}^n \widehat{\nu}_{2i}^2 = \sum_{i=1}^n \widehat{\nu}_{2i}\widetilde{\omega}_i / \sum_{i=1}^n \widehat{\nu}_{2i}^2 \tag{A.11}$$

where the last equality follows from the fact that $\widetilde{Y}$ is a linear combination of all X's excluding X_2, all of which satisfy (A.1). Hence $\widehat{\beta}_2$ is the estimate of the slope coefficient in the linear regression of $\widetilde{\omega}$ on $\widehat{\nu}_2$.

Simple, Partial and Multiple Correlation Coefficients

In Chapter 3, we interpreted the square of the *simple correlation coefficient*, r_{Y,X_2}^2, as the proportion of the variation in Y that is explained by X_2. Similarly, r_{Y,X_k}^2 is the R-squared of the simple regression of Y on X_k for $k = 2, \ldots, K$. In fact, one can compute these simple correlation coefficients and find out which X_k is most correlated with Y, say it is X_2. If one is selecting regressors to include in the regression equation, X_2 would be the best one variable candidate. In order to determine what variable to include next, we look at *partial correlation coefficients* of the form $r_{Y,X_k.X_2}$ for $k \neq 2$. The square of this first-order partial gives the proportion of the residual variation in Y, not explained by X_2, that is explained by the addition of X_k. The maximum first-order partial ('first' because it has only one variable after the dot) determines the best candidate to follow X_2. Let us assume it is X_3. The first-order partial correlation coefficients can be computed from simple correlation coefficients as follows:

$$r_{Y,X_3.X_2} = \frac{r_{Y,X_3} - r_{Y,X_2}r_{X_2,X_3}}{\sqrt{1 - r_{Y,X_2}^2}\sqrt{1 - r_{X_2,X_3}^2}}$$

see Johnston (1984). Next we look at second-order partials of the form $r_{Y,X_k.X_2,X_3}$ for $k \neq 2, 3$, and so on. This method of selecting regressors is called *forward selection*. Suppose there is only X_2, X_3 and X_4 in the regression equation. In this case $(1 - r_{Y,X_2}^2)$ is the proportion of the variation in Y, i.e., $\sum_{i=1}^n y_i^2$, that is not explained by X_2. Also $(1 - r_{Y,X_3.X_2}^2)(1 - r_{Y,X_2}^2)$ denotes the proportion of the variation in Y not explained after the inclusion of both X_2 and X_3. Similarly $(1 - r_{Y,X_4.X_2,X_3}^2)(1 - r_{Y,X_3.X_2}^2)(1 - r_{Y,X_2}^2)$ is the proportion of the variation in Y unexplained after the inclusion of X_2, X_3 and X_4. But this is exactly $(1 - R^2)$, where R^2 denotes the R-squared of the multiple regression of Y on a constant, X_2, X_3 and X_4. This R^2 is called the *multiple correlation coefficient*, and is also written as $R_{Y.X_2,X_3,X_4}^2$. Hence

$$(1 - R_{Y.X_2,X_3,X_4}^2) = (1 - r_{Y,X_2}^2)(1 - r_{Y,X_3.X_2}^2)(1 - r_{Y,X_4.X_2,X_3}^2)$$

and similar expressions relating the multiple correlation coefficient to simple and partial correlation coefficients can be written by including say X_3 first then X_4 and X_2 in that order.

CHAPTER 5
Violations of the Classical Assumptions

5.1 Introduction

In this chapter, we relax the assumptions made in Chapter 3 one by one and study the effect of that on the OLS estimator. In case the OLS estimator is no longer a viable estimator, we derive an alternative estimator and propose some tests that will allow us to check whether this assumption is violated.

5.2 The Zero Mean Assumption

Violation of assumption 1 implies that the mean of the disturbances is no longer zero. Two cases are considered:

Case 1: $E(u_i) = \mu \neq 0$

The disturbances have a common mean which is not zero. In this case, one can subtract μ from the u_i's and get new disturbances $u_i^* = u_i - \mu$ which have zero mean and satisfy all the other assumptions imposed on the u_i's. Having subtracted μ from u_i we add it to the constant α leaving the regression equation intact:

$$Y_i = \alpha^* + \beta X_i + u_i^* \quad i = 1, 2, \ldots, n \tag{5.1}$$

where $\alpha^* = \alpha + \mu$. It is clear that only α^* and β can be estimated, and not α nor μ. In other words, one cannot retrieve α and μ from an estimate of α^* without additional assumptions or further information, see problem 10. With this reparameterization, equation (5.1) satisfies the four classical assumptions, and therefore OLS gives the BLUE estimators of α^* and β. Hence, a constant non-zero mean for the disturbances affects only the intercept estimate but not the slope. Fortunately, in most economic applications, it is the slope coefficients that are of interest and not the intercept.

Case 2: $E(u_i) = \mu_i$

The disturbances have a mean which varies with every observation. In this case, one can transform the regression equation as in (5.1) by adding and subtracting μ_i. The problem, however, is that $\alpha_i^* = \alpha + \mu_i$ now varies with each observation, and hence we have more parameters than observations. In fact, there are n intercepts and one slope to be estimated with n observations. Unless we have repeated observations like in panel data, see Chapter 12 or we have some prior information on these α_i^*, we cannot estimate this model.

5.3 Stochastic Explanatory Variables

Sections 5.5 and 5.6 will study violations of assumptions 2 and 3 in detail. This section deals with violations of assumption 4 and its effect on the properties of the OLS estimators. In this case, X is a random variable which may be (i) independent; (ii) contemporaneously uncorrelated; or (iii) simply correlated with the disturbances.

Case 1: If X is independent of u, then all the results of Chapter 3 still hold, but now they are conditional on the particular set of X's drawn in the sample. To illustrate this result, recall that for the simple linear regression:

$$\widehat{\beta}_{OLS} = \beta + \textstyle\sum_{i=1}^{n} w_i u_i \text{ where } w_i = x_i / \sum_{i=1}^{n} x_i^2 \tag{5.2}$$

Hence, when we take expectations $E(\sum_{i=1}^{n} w_i u_i) = \sum_{i=1}^{n} E(w_i)E(u_i) = 0$. The first equality holds because X and u are independent and the second equality holds because the u's have zero mean. In other words the unbiasedness property of the OLS estimator still holds. However, the

$$\text{var}(\widehat{\beta}_{OLS}) = E(\textstyle\sum_{i=1}^{n} w_i u_i)^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} E(w_i w_j) E(u_i u_j) = \sigma^2 \sum_{i=1}^{n} E(w_i^2)$$

where the last equality follows from assumptions 2 and 3, homoskedasticity and no serial correlation. The only difference between this result and that of Chapter 3 is that we have expectations on the X's rather than the X's themselves. Hence, by conditioning on the particular set of X's that are observed, we can use all the results of Chapter 3. Also, maximizing the likelihood involves both the X's and the u's. But, as long as the distribution of the X's does not involve the parameters we are estimating, i.e., α, β and σ^2, the same maximum likelihood estimators are obtained. Why? Because $f(x_1, x_2, \ldots, x_n, u_1, u_2, \ldots, u_n) = f_1(x_1, x_2, \ldots, x_n) f_2(u_1, u_2, \ldots, u_n)$ since the X's and the u's are independent. Maximizing f with respect to $(\alpha, \beta, \sigma^2)$ is the same as maximizing f_2 with respect to $(\alpha, \beta, \sigma^2)$ as long as f_1 is not a function of these parameters.

Case 2: Consider a simple model of consumption, where Y_t, current consumption, is a function of Y_{t-1}, consumption in the previous period. This is the case for a habit forming consumption good like cigarette smoking. In this case our regression equation becomes

$$Y_t = \alpha + \beta Y_{t-1} + u_t \quad t = 2, \ldots, T \tag{5.3}$$

where we lost one observation due to lagging. It is obvious that Y_t is correlated to u_t, but the question here is whether Y_{t-1} is correlated to u_t. After all, Y_{t-1} is our explanatory variable X_t. As long as assumption 3 is not violated, i.e., the u's are not correlated across periods, u_t represents a freshly drawn disturbance independent of previous disturbances and hence is not correlated with the already *predetermined* Y_{t-1}. This is what we mean by contemporaneously uncorrelated, i.e., u_t is correlated with Y_t, but it is not correlated with Y_{t-1}. The OLS estimator of β is

$$\widehat{\beta}_{OLS} = \textstyle\sum_{t=2}^{T} y_t y_{t-1} / \sum_{t=2}^{T} y_{t-1}^2 = \beta + \sum_{t=2}^{T} y_{t-1} u_t / \sum_{t=2}^{T} y_{t-1}^2 \tag{5.4}$$

and the expected value of (5.4) is not β because in general,

$$E(\textstyle\sum_{t=2}^{T} y_{t-1} u_t / \sum_{t=2}^{T} y_{t-1}^2) \neq E(\sum_{t=2}^{T} y_{t-1} u_t) / E(\sum_{t=2}^{T} y_{t-1}^2).$$

The expected value of a ratio is not the ratio of expected values. Also, even if $E(Y_{t-1} u_t) = 0$, one can easily show that $E(y_{t-1} u_t) \neq 0$. In fact, $y_{t-1} = Y_{t-1} - \bar{Y}$, and $\bar{Y}$ contains Y_t in it, and

we know that $E(Y_t u_t) \neq 0$. Hence, we lost the unbiasedness property of OLS. However, all the asymptotic properties still hold. In fact, $\widehat{\beta}_{OLS}$ is consistent because

$$\text{plim } \widehat{\beta}_{OLS} = \beta + \text{cov}(Y_{t-1}, u_t)/\text{var}(Y_{t-1}) = \beta \tag{5.5}$$

where the second equality follows from (5.4) and the fact that $\text{plim}(\sum_{t=2}^{T} y_{t-1}u_t/T)$ is $\text{cov}(Y_{t-1}, u_t)$ which is zero, and $\text{plim}(\sum_{t=2}^{T} y_{t-1}^2/T) = \text{var}(Y_{t-1})$ which is positive and finite.

Case 3: X and u are correlated, in this case OLS is biased and inconsistent. This can be easily deduced from (5.2) since $\text{plim}(\sum_{i=1}^{n} x_i u_i/n)$ is the $\text{cov}(X, u) \neq 0$, and $\text{plim}(\sum_{i=1}^{n} x_i^2/n)$ is positive and finite. This means that OLS is no longer a viable estimator, and an alternative estimator that corrects for this bias has to be derived. In fact we will study three specific cases where this assumption is violated. These are: (i) the errors in measurement case; (ii) the case of a lagged dependent variable with correlated errors; and (iii) simultaneous equations.

Briefly, the errors in measurement case involves a situation where the true regression model is in terms of X^*, but X^* is measured with error, i.e., $X_i = X_i^* + \nu_i$, so we observe X_i but not X_i^*. Hence, when we substitute this X_i for X_i^* in the regression equation, we get

$$Y_i = \alpha + \beta X_i^* + u_i = \alpha + \beta X_i + (u_i - \beta\nu_i) \tag{5.6}$$

where the composite error term is now correlated with X_i because X_i is correlated with ν_i. After all, $X_i = X_i^* + \nu_i$ and $E(X_i\nu_i) = E(\nu_i^2)$ if X_i^* and ν_i are uncorrelated.

Similarly, in case (ii) above, if the u's were correlated across time, i.e., u_{t-1} is correlated with u_t, then Y_{t-1}, which is a function of u_{t-1}, will also be correlated with u_t, and $E(Y_{t-1}u_t) \neq 0$. More on this and how to test for serial correlation in the presence of a lagged dependent variable in Chapter 6.

Finally, if one considers a demand and supply equations where quantity Q_t is a function of price P_t in both equations

$$Q_t = \alpha + \beta P_t + u_t \quad \text{(demand)} \tag{5.7}$$

$$Q_t = \delta + \gamma P_t + \nu_t \quad \text{(supply)} \tag{5.8}$$

The question here is whether P_t is correlated with the disturbances u_t and ν_t in both equations. The answer is yes, because (5.7) and (5.8) are two equations in two unknowns P_t and Q_t. Solving for these variables, one gets P_t as well as Q_t as a function of a constant and both u_t and ν_t. This means that $E(P_t u_t) \neq 0$ and $E(P_t\nu_t) \neq 0$ and OLS performed on either (5.7) or (5.8) is biased and inconsistent. We will study this simultaneous bias problem more rigorously in Chapter 11.

For all situations where X and u are correlated, it would be illuminating to show graphically why OLS is no longer a consistent estimator. Let us consider the case where the disturbances are, say, positively correlated with the explanatory variable. Figure 3.3 of Chapter 3 shows the true regression line $\alpha + \beta X_i$. It also shows that when X_i and u_i are positively correlated then an X_i higher than its mean will be associated with a disturbance u_i above its mean, i.e., a positive disturbance. Hence, $Y_i = \alpha + \beta X_i + u_i$ will always be above the true regression line whenever X_i is above its mean. Similarly Y_i would be below the true regression line for every X_i below its mean. This means that not knowing the true regression line, a researcher fitting OLS on this data will have a biased intercept and slope. In fact, the intercept will be understated and the slope will be overstated. Furthermore, this bias does not disappear with more data, since

this new data will be generated by the same mechanism described above. Hence these OLS estimates are inconsistent.

Similarly, if X_i and u_i are negatively correlated, the intercept will be overstated and the slope will be understated. This story applies to any equation with at least one of its right hand side variables correlated with the disturbance term. Correlation due to the lagged dependent variable with autocorrelated errors, is studied in Chapter 6, whereas the correlation due to the simultaneous equations problem is studied in Chapter 11.

5.4 Normality of the Disturbances

If the disturbance are not normal, OLS is still BLUE provided assumptions 1-4 still hold. Normality made the OLS estimators minimum variance unbiased MVU and these OLS estimators turn out to be identical to the MLE. Normality allowed the derivation of the distribution of these estimators and this in turn allowed testing of hypotheses using the t and F-tests considered in the previous chapter. If the disturbances are not normal, yet the sample size is large, one can still use the normal distribution for the OLS estimates asymptotically by relying on the Central Limit Theorem, see Theil (1978). Theil's proof is for the case of fixed X's in repeated samples, zero mean and constant variance on the disturbances. A simple asymptotic test for the normality assumption is given by Jarque and Bera (1987). This is based on the fact that the normal distribution has a skewness measure of zero and a kurtosis of 3. Skewness (or lack of symmetry) is measured by

$$S = \frac{[E(X-\mu)^3]^2}{[E(X-\mu)^2]^3} = \frac{Square\ of\ the\ 3rd\ moment\ about\ the\ mean}{Cube\ of\ the\ variance}$$

Kurtosis (a measure of flatness) is measured by

$$\kappa = \frac{E(X-\mu)^4}{[E(X-\mu)^2]^2} = \frac{4th\ moment\ about\ the\ mean}{Square\ of\ the\ variance}$$

For the normal distribution $S = 0$ and $\kappa = 3$. Hence, the Jarque-Bera (JB) statistic is given by

$$JB = n\left[\frac{S^2}{6} + \frac{(\kappa-3)^2}{24}\right]$$

where S represents skewness and κ represents kurtosis of the OLS residuals. This statistic is asymptotically distributed as χ^2 with two degrees of freedom under H_0. Rejecting H_0, rejects normality of the disturbances but does not offer an alternative distribution. In this sense, the test is non-constructive. In addition, not rejecting H_0 does not necessarily mean that the distribution of the disturbances is normal, it only means we do not reject that the distribution of the disturbances is symmetric and has a kurtosis of 3. See the empirical example in section 5.5 for an illustration. The Jarque-Bera test is part of the standard output using EViews.

5.5 Heteroskedasticity

Violation of assumption 2, means that the disturbances have a varying variance, i.e., $E(u_i^2) = \sigma_i^2$, $i = 1, 2, \ldots, n$. First, we study the effect of this violation on the OLS estimators. For the simple

linear regression it is obvious that $\widehat{\beta}_{OLS}$ given in equation (5.2) is still unbiased and consistent because these properties depend upon assumptions 1 and 4, and not assumption 2. However, the variance of $\widehat{\beta}_{OLS}$ is now different

$$\text{var}(\widehat{\beta}_{OLS}) = \text{var}(\textstyle\sum_{i=1}^{n} w_i u_i) = \sum_{i=1}^{n} w_i^2 \sigma_i^2 = \sum_{i=1}^{n} x_i^2 \sigma_i^2 / (\sum_{i=1}^{n} x_i^2)^2 \tag{5.9}$$

where the second equality follows from assumption 3 and the fact that $\text{var}(\text{u}_i)$ is now σ_i^2. Note that if $\sigma_i^2 = \sigma^2$, this reverts back to $\sigma^2 / \sum_{i=1}^{n} x_i^2$, the usual formula for $\text{var}(\widehat{\beta}_{OLS})$ under homoskedasticity. Furthermore, one can show that $E(s^2)$ will involve all of the σ_i^2's and not one common σ^2, see problem 1. This means that the regression package reporting $s^2 / \sum_{i=1}^{n} x_i^2$ as the estimate of the variance of $\widehat{\beta}_{OLS}$ is committing two errors. One, it is not using the right formula for the variance, i.e., equation (5.9). Second, it is using s^2 to estimate a common σ^2 when in fact the σ_i^2's are different. The bias from using $s^2 / \sum_{i=1}^{n} x_i^2$ as an estimate of $\text{var}(\widehat{\beta}_{OLS})$ will depend upon the nature of the heteroskedasticity and the regressor. In fact, if σ_i^2 is positively related to x_i^2, one can show that $s^2 / \sum_{i=1}^{n} x_i^2$ understates the true variance and hence the t-statistic reported for $\beta = 0$ is overblown, and the confidence interval for β is tighter than it is supposed to be, see problem 2. This means that the t-statistic in this case is biased towards rejecting $H_0; \beta = 0$, i.e., showing significance of the regression slope coefficient, when it may not be significant.

The OLS estimator of β is linear unbiased and consistent, but is it still BLUE? In order to answer this question, we note that the only violation we have is that the $\text{var}(u_i) = \sigma_i^2$. Hence, if we divided u_i by σ_i / σ, the resulting $u_i^* = \sigma u_i / \sigma_i$ will have a constant variance σ^2. It is easy to show that u^* satisfies all the classical assumptions including homoskedasticity. The regression model becomes

$$\sigma Y_i / \sigma_i = \alpha \sigma / \sigma_i + \beta \sigma X_i / \sigma_i + u_i^* \tag{5.10}$$

and OLS on this model (5.10) is BLUE. The OLS normal equations on (5.10) are

$$\begin{aligned} \textstyle\sum_{i=1}^{n} (Y_i / \sigma_i^2) &= \alpha \textstyle\sum_{i=1}^{n} (1/\sigma_i^2) + \beta \sum_{i=1}^{n} (X_i / \sigma_i^2) \\ \textstyle\sum_{i=1}^{n} (Y_i X_i / \sigma_i^2) &= \alpha \textstyle\sum_{i=1}^{n} (X_i / \sigma_i^2) + \beta \sum_{i=1}^{n} (X_i^2 / \sigma_i^2) \end{aligned} \tag{5.11}$$

Note that σ^2 drops out of these equations. Solving (5.11), see problem 3, one gets

$$\widetilde{\alpha} = [\textstyle\sum_{i=1}^{n} (Y_{1i} / \sigma_i^2) / \sum_{i=1}^{n} (1/\sigma_i^2)] - \widetilde{\beta} [\sum_{i=1}^{n} (X_i / \sigma_i^2) / \sum_{i=1}^{n} (1/\sigma_i^2)] = \bar{Y}^* - \widetilde{\beta} \bar{X}^* \tag{5.12a}$$

with $\bar{Y}^* = [\sum_{i=1}^{n} (Y_i / \sigma_i^2) / \sum_{i=1}^{n} (1/\sigma_i^2)] = \sum_{i=1}^{n} w_i^* Y_i / \sum_{i=1}^{n} w_i^*$ and

$$\bar{X}^* = [\textstyle\sum_{i=1}^{n} (X_i / \sigma_i^2) / \sum_{i=1}^{n} (1/\sigma_i^2)] = \sum_{i=1}^{n} w_i^* X_i / \sum_{i=1}^{n} w_i^*$$

where $w_i^* = (1/\sigma_i^2)$. Similarly,

$$\begin{aligned} \widetilde{\beta} &= \frac{[\sum_{i=1}^{n} (1/\sigma_i^2)][\sum_{i=1}^{n} (Y_i X_i / \sigma_i^2)] - [\sum_{i=1}^{n} (X_i / \sigma_i^2)][\sum_{i=1}^{n} (Y_i / \sigma_i^2)]}{[\sum_{i=1}^{n} X_i^2 / \sigma_i^2][\sum_{i=1}^{n} (1/\sigma_i^2)] - [\sum_{i=1}^{n} (X_i / \sigma_i^2)]^2} \\ &= \frac{(\sum_{i=1}^{n} w_i^*)(\sum_{i=1}^{n} w_i^* X_i Y_i) - (\sum_{i=1}^{n} w_i^* X_i)(\sum_{i=1}^{n} w_i^* Y_i)}{(\sum_{i=1}^{n} w_i^*)(\sum_{i=1}^{n} w_i^* X_i^2) - (\sum_{i=1}^{n} w_i^* X_i)^2} \\ &= \frac{\sum_{i=1}^{n} w_i^* (X_i - \bar{X}^*)(Y_i - \bar{Y}^*)}{\sum_{i=1}^{n} w_i^* (X_i - \bar{X}^*)^2} \end{aligned} \tag{5.12b}$$

It is clear that the BLU estimators $\widetilde{\alpha}$ and $\widetilde{\beta}$, obtained from the regression in (5.10), are different from the usual OLS estimators $\widehat{\alpha}_{OLS}$ and $\widehat{\beta}_{OLS}$ since they depend upon the σ_i^2's. It is also true that when $\sigma_i^2 = \sigma^2$ for all $i = 1, 2, \ldots, n$, i.e., under homoskedasticity, (5.12) reduces to the usual OLS estimators given by equation (3.4) of Chapter 3. The BLU estimators weight the i-th observation by $(1/\sigma_i)$ which is a measure of precision of that observation. The more precise the observation, i.e., the smaller σ_i, the larger is the weight attached to that observation. $\widetilde{\alpha}$ and $\widetilde{\beta}$ are also known as *Weighted Least Squares* (WLS) estimators which are a specific form of *Generalized Least Squares* (GLS). We will study GLS in details in Chapter 9, using matrix notation.

Under heteroskedasticity, OLS looses efficiency in that it is no longer BLUE. However, because it is still unbiased and consistent and because the true σ_i^2's are never known some researchers compute OLS as an initial consistent estimator of the regression coefficients. It is important to emphasize however, that the standard errors of these estimates as reported by the regression package are biased and any inference based on these estimated variances including the reported t-statistics are misleading. White (1980) proposed a simple procedure that would yield heteroskedasticity consistent standard errors of the OLS estimators. In equation (5.9), this amounts to replacing σ_i^2 by e_i^2, the square of the i-th OLS residual, i.e.,

$$White's\ \text{var}(\widehat{\beta}_{OLS}) = \textstyle\sum_{i=1}^{n} x_i^2 e_i^2 / (\sum_{i=1}^{n} x_i^2)^2 \tag{5.13}$$

Note that we can not consistently estimate σ_i^2 by e_i^2, since there is one observation per parameter estimated. As the sample size increases, so does the number of unknown σ_i^2's. What White (1980) consistently estimates is the $\text{var}(\widehat{\beta}_{OLS})$ which is a weighted average of the e_i^2. The same analysis applies to the multiple regression OLS estimates. In this case, White's (1980) heteroskedasticity consistent estimate of the variance of the k-th OLS regression coefficient β_k, is given by

$$White's\ \text{var}(\widehat{\beta}_k) = \textstyle\sum_{i=1}^{n} \widehat{\nu}_{ki}^2 e_i^2 / (\sum_{i=1}^{n} \widehat{\nu}_{ki}^2)^2$$

where $\widehat{\nu}_k^2$ is the squared OLS residual obtained from regressing X_k on the remaining regressors in the equation being estimated. e_i is the i-th OLS residual from this multiple regression equation. Many regression packages provide White's heteroskedasticity-consistent estimates of the variances and their corresponding robust t-statistics. For example, using EViews, one clicks on Quick, choose Estimate Equation. Now click on Options, a menu appears where one selects White to obtain the heteroskedasticity-consistent estimates of the variances.

While the regression packages correct for heteroskedasticity in the t-statistics they do not usually do that for the F-statistics studied, say in Example 2 in Chapter 4. Wooldridge (1991) suggests a simple way of obtaining a robust LM statistic for H_0; $\beta_2 = \beta_3 = 0$ in the multiple regression (4.1). This involves the following steps:

(1) Run OLS on the restricted model without X_2 and X_3 and obtain the restricted least squares residuals $\widetilde{u}$.

(2) Regress each of the independent variables excluded under the null (i.e., X_2 and X_3) on *all* of the other included independent variables (i.e., $X_4, X_5, \ldots, X_K$) including the constant. Get the corresponding residuals $\widehat{v}_2$ and $\widehat{v}_3$, respectively.

(3) Regress the dependent variable equal to 1 for all observations on $\widehat{v}_2\widetilde{u}, \widehat{v}_3\widetilde{u}$ without a constant and obtain the robust LM statistic equal to the n- the sum of squared residuals of

this regression. This is exactly nR_u^2 of this last regression. Under H_0 this LM statistic is distributed as χ_2^2.

Since OLS is no longer BLUE, one should compute $\widetilde{\alpha}$ and $\widetilde{\beta}$. The only problem is that the σ_i's are rarely known. One example where the σ_i's are known up to a scalar constant is the following simple example of aggregation.

Example 5.1: *Aggregation and Heteroskedasticity.* Let Y_{ij} be the observation on the j-th firm in the i-th industry, and consider the following regression:

$$Y_{ij} = \alpha + \beta X_{ij} + u_{ij} \quad j = 1, 2, \ldots, n_i;\ i = 1, 2, \ldots, m \tag{5.14}$$

If only aggregate observations on each industry are available, then (5.14) is summed over firms, i.e.,

$$Y_i = \alpha n_i + \beta X_i + u_i \quad i = 1, 2, \ldots, m \tag{5.15}$$

where $Y_i = \sum_{j=1}^{n_i} Y_{ij}$, $X_i = \sum_{j=1}^{n_i} X_{ij}$, $u_i = \sum_{j=1}^{n_i} u_{ij}$ for $i = 1, 2, \ldots, m$. Note that although the u_{ij}'s are IID$(0, \sigma^2)$, by aggregating, we get $u_i \sim (0, n_i\sigma^2)$. This means that the disturbances in (5.15) are heteroskedastic. However, $\sigma_i^2 = n_i\sigma^2$ and is known up to a scalar constant. In fact, σ/σ_i is $1/(n_i)^{1/2}$. Therefore, premultiplying (5.15) by $1/(n_i)^{1/2}$ and performing OLS on the transformed equation results in BLU estimators of α and β. In other words, BLU estimation reduces to performing OLS of $Y_i/(n_i)^{1/2}$ on $(n_i)^{1/2}$ and $X_i/(n_i)^{1/2}$, without an intercept.

There may be other special cases in practice where σ_i is known up to a scalar, but in general, σ_i is usually unknown and will have to be estimated. This is hopeless with only n observations, since there are n σ_i's, so we either have to have repeated observations, or know more about the σ_i's. Let us discuss these two cases.

Case 1: *Repeated Observations*

Suppose that n_i households are selected randomly with income X_i for $i = 1, 2, \ldots, m$. For each household $j = 1, 2, \ldots, n_i$, we observe its consumption expenditures on food, say Y_{ij}. The regression equation is

$$Y_{ij} = \alpha + \beta X_i + u_{ij} \quad i = 1, 2, \ldots, m\ ;\ j = 1, 2, \ldots, n_i \tag{5.16}$$

where m is the number of income groups selected. Note that X_i has only one subscript, whereas Y_{ij} has two subscripts denoting the repeated observations on households with the same income X_i. The u_{ij}'s are independently distributed $(0, \sigma_i^2)$ reflecting the heteroskedasticity in consumption expenditures among the different income groups. In this case, there are $n = \sum_{i=1}^{m} n_i$ observations and m σ_i^2's to be estimated. This is feasible, and there are two methods for estimating these σ_i^2's. The first is to compute

$$\widehat{s}_i^2 = \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2/(n_i - 1)$$

where $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$. The second is to compute $\widetilde{s}_i^2 = \sum_{j=1}^{n_i} e_{ij}^2/n_i$ where e_{ij} is the OLS residual given by

$$e_{ij} = Y_{ij} - \widehat{\alpha}_{OLS} - \widehat{\beta}_{OLS} X_i$$

Both estimators of σ_i^2 are consistent. Substituting either $\widetilde{s}_i^2$ or $\widehat{s}_i^2$ for σ_i^2 in (5.12) will result in feasible estimators of α and β. However, the resulting estimates are no longer BLUE. The substitution of the consistent estimators of σ_i^2 is justified on the basis that the resulting α and β estimates will be asymptotically efficient, see Chapter 9. Of course, this step could have been replaced by a regression of $Y_{ij}/\widehat{s}_i$ on $(1/\widehat{s}_i)$ and $(X_i/\widehat{s}_i)$ without a constant, or the similar regression in terms of $\widetilde{s}_i$. For this latter estimate, $\widetilde{s}_i^2$, one can iterate, i.e., obtaining new residuals based on the new regression estimates and therefore new $\widetilde{s}_i^2$. The process continues until the estimates obtained from the r-th iteration do not differ from those of the $(r+1)$th iteration in absolute value by more than a small arbitrary positive number chosen as the convergence criterion. Once the estimates converge, the final round estimators are the maximum likelihood estimators, see Oberhofer and Kmenta (1974).

Case 2: *Assuming More Information on the Form of Heteroskedasticity*

If we do not have repeated observations, it is hopeless to try and estimate n variances and α and β with only n observations. More structure on the form of heteroskedasticity is needed to estimate this model, but not necessarily to test it. Heteroskedasticity is more likely to occur with cross-section data where the observations may be on firms with different size. For example, a regression relating profits to sales might have heteroskedasticity, because larger firms have more resources to draw upon, can borrow more, invest more, and loose or gain more than smaller firms. Therefore, we expect the form of heteroskedasticity to be related to the size of the firm, which is reflected in this case by the regressor, sales, or some other variable that measures size, like assets. Hence, for this regression we can write $\sigma_i^2 = \sigma^2 Z_i^2$, where Z_i denotes the sales or assets of firm i. Once again the form of heteroskedasticity is known up to a scalar constant and the BLU estimators of α and β can be obtained from (5.12), assuming Z_i is known. Alternatively, one can run the regression of Y_i/Z_i on $1/Z_i$ and X_i/Z_i without a constant to get the same result. Special cases of Z_i are X_i and $E(Y_i)$. (i) If $Z_i = X_i$ the regression becomes that of Y_i/X_i on $1/X_i$ and a constant. Note that the regression coefficient of $1/X_i$ is the estimate of α, while the constant of the regression is now the estimate of β. But, is it possible to have u_i uncorrelated with X_i when we are assuming $\text{var}(u_i)$ related to X_i? The answer is yes, as long as $E(u_i/X_i) = 0$, i.e., the mean of u_i is zero for every value of X_i, see Figure 3.4 of Chapter 3. This, in turn, implies that the overall mean of the u_i's is zero, i.e., $E(u_i) = 0$ and that $\text{cov}(X_i, u_i) = 0$. If the latter is not satisfied and say $\text{cov}(X_i, u_i)$ is positive, then large values of X_i imply large values of u_i. This would mean that for these values of X_i, we have a non-zero mean for the corresponding u_i's. This contradicts $E(u_i/X_i) = 0$. Hence, if $E(u_i/X_i) = 0$, then $\text{cov}(X_i, u_i) = 0$. (ii) If $Z_i = E(Y_i) = \alpha + \beta X_i$, then σ_i^2 is proportional to the population regression line, which is a linear function of α and β. Since the OLS estimates are consistent one can estimate $E(Y_i)$ by $\widehat{Y}_i = \widehat{\alpha}_{OLS} + \widehat{\beta}_{OLS} X_i$ use $\widehat{Z}_i = \widehat{Y}_i$ instead of $E(Y_i)$. In other words, run the regression of $Y_i/\widehat{Y}_i$ on $1/\widehat{Y}_i$ and $X_i/\widehat{Y}_i$ without a constant. The resulting estimates are asymptotically efficient, see Amemiya (1973).

One can generalize $\sigma_i^2 = \sigma^2 Z_i^2$ to $\sigma_i^2 = \sigma^2 Z_i^\delta$ where δ is an unknown parameter to be estimated. Hence rather than estimating n σ_i^2's one has to estimate only σ^2 and δ. Assuming normality one can set up the likelihood function and derive the first-order conditions by differentiating that likelihood with respect to α, β, σ^2 and δ. The resulting equations are highly nonlinear. Alternatively, one can search over possible values for $\delta = 0, 0.1, 0.2, \ldots, 4$, and get the corresponding estimates of α, β, and σ^2 from the regression of $Y_i/Z_i^{\delta/2}$ on $1/Z_i^{\delta/2}$ and $X_i/Z_i^{\delta/2}$

without a constant. This is done for every δ and the value of the likelihood function is reported. Using this search procedure one can get the maximum value of the likelihood and corresponding to it the MLE of α, β, σ^2 and δ. Note that as δ increases so does the degree of heteroskedasticity. Problem 4 asks the reader to compute the relative efficiency of the OLS estimator with respect to the BLU estimator for $Z_i = X_i$ for various values of δ. As expected the relative efficiency of the OLS estimator declines as the degree of heteroskedasticity increases.

One can also generalize $\sigma_i^2 = \sigma^2 Z_i^\delta$ to include more Z variables. In fact, a general form of this multiplicative heteroskedasticity is

$$\log\sigma_i^2 = \log\sigma^2 + \delta_1 \log Z_{1i} + \delta_2 \log Z_{2i} + \ldots + \delta_r \log Z_{ri} \tag{5.17}$$

with $r < n$, otherwise one cannot estimate with n observations. $Z_1, Z_2, \ldots, Z_r$ are known variables determining the heteroskedasticity. Note that if $\delta_2 = \delta_3 = \ldots = \delta_r = 0$, we revert back to $\sigma_i^2 = \sigma^2 Z_i^\delta$, where $\delta = \delta_1$. For the estimation of this general multiplicative form of heteroskedasticity, see Harvey (1976).

Another form for heteroskedasticity, is the additive form

$$\sigma_i^2 = a + b_1 Z_{1i} + b_2 Z_{2i} + \ldots + b_r Z_{ri} \tag{5.18}$$

where $r < n$, see Goldfeld and Quandt (1972). Special cases of (5.18) include

$$\sigma_i^2 = a + b_1 X_i + b_2 X_i^2 \tag{5.19}$$

where if a and b_1 are zero we have a simple form of multiplicative heteroskedasticity. In order to estimate the regression model with additive heteroskedasticity of the type given in (5.19), one can get the OLS residuals, the e_i's, and run the following regression

$$e_i^2 = a + b_1 X_i + b_2 X_i^2 + v_i \tag{5.20}$$

where $v_i = e_i^2 - \sigma_i^2$. The v_i's are heteroskedastic, and the OLS estimates of (5.20) yield the following estimates of σ_i^2

$$\widehat{\sigma}_i^2 = \widehat{a}_{OLS} + \widehat{b}_{1,OLS} X_i + \widehat{b}_{2,OLS} X_i^2 \tag{5.21}$$

One can obtain a better estimate of the σ_i^2's by computing the following regression which corrects for the heteroskedasticity in the v_i's

$$(e_i^2/\widehat{\sigma}_i) = a(1/\widehat{\sigma}_i) + b_1(X_i/\widehat{\sigma}_i) + b_2(X_i^2/\widehat{\sigma}_i) + w_i \tag{5.22}$$

The new estimates of σ_i^2 are

$$\widetilde{\sigma}_i^2 = \widetilde{a} + \widetilde{b}_1 X_i + \widetilde{b}_2 X_i^2 \tag{5.23}$$

where $\widetilde{a}$, $\widetilde{b}_1$ and $\widetilde{b}_2$ are the OLS estimates from (5.22). Using the $\widetilde{\sigma}_i^2$'s one can run the regression of $Y_i/\widetilde{\sigma}_i$ on $(1/\widetilde{\sigma}_i)$ and $X_i/\widetilde{\sigma}_i$ without a constant to get asymptotically efficient estimates of α and β. These have the same asymptotic properties as the MLE estimators derived in Rutemiller and Bowers (1968), see Amemiya (1977) and Buse (1984). The problem with this iterative procedure is that there is no guarantee that the $\widetilde{\sigma}_i^2$'s are positive, which means that the square root $\widetilde{\sigma}_i$ may not exist. This problem would not occur if $\sigma_i^2 = (a + b_1 X_i + b_2 X_i^2)^2$ because in this case one regresses $|e_i|$ on a constant, X_i and X_i^2 and the predicted value from this regression would be an estimate of σ_i. It would not matter if this predictor is negative, because we do not have to take its square root and because its sign cancels in the OLS normal equations of the final regression of $Y_i/\widehat{\sigma}_i$ on $(1/\widehat{\sigma}_i)$ and $(X_i/\widehat{\sigma}_i)$ without a constant.

Testing for Homoskedasticity

In the repeated observation's case, one can perform Bartlett's (1937) test. The null hypothesis is $H_0; \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_m^2$. Under the null there is one variance σ^2 which can be estimated by the pooled variance $s^2 = \sum_{i=1}^{m} \nu_i \widetilde{s}_i^2/\nu$ where $\nu = \sum_{i=1}^{m} \nu_i$, and $\nu_i = n_i - 1$. Under the alternative hypothesis there are m different variances estimated by $\widetilde{s}_i^2$ for $i = 1, 2, \ldots, m$. The Likelihood Ratio test, which computes the ratio of the likelihoods under the null and alternative hypotheses, reduces to computing

$$B = [\nu \log s^2 - \textstyle\sum_{i=1}^{m} \nu_i \log s_i^2]/c \tag{5.24}$$

where $c = 1 + [\sum_{i=1}^{m}(1/\nu_i) - 1/\nu]/3(m-1)$. Under H_0, B is distributed χ^2_{m-1}. Hence, a large p-value for the B-statistic given in (5.24) means that we do not reject homoskedasticity whereas, a small p-value leads to rejection of H_0 in favor of heteroskedasticity.

In case of no repeated observations, several tests exist in the literature. Among these are the following:

(1) Glejser's (1969) Test: In this case one regresses $|e_i|$ on a constant and Z_i^{δ} for $\delta = 1, -1, 0.5$ and -0.5. If the coefficient of Z_i^{δ} is significantly different from zero, this would lead to a rejection of homoskedasticity. The power of this test depends upon the true form of heteroskedasticity. One important result however, is that this power is not seriously impaired if the wrong value of δ is chosen, see Ali and Giaccotto (1984) who confirmed this result using extensive Monte Carlo experiments.

(2) The Goldfeld and Quandt (1965) Test: This is a simple and intuitive test. One orders the observations according to X_i and omits c central observations. Next, two regressions are run on the two separated sets of observations with $(n-c)/2$ observations in each. The c omitted observations separate the low value X's from the high value X's, and if heteroskedasticity exists and is related to X_i, the estimates of σ^2 reported from the two regressions should be different. Hence, the test statistic is s_2^2/s_1^2 where s_1^2 and s_2^2 are the Mean Square Error of the two regressions, respectively. Their ratio would be the same as that of the two residual sums of squares because the degrees of freedom of the two regressions are the same. This statistic is F-distributed with $((n-c)/2) - K$ degrees of freedom in the numerator as well as the denominator. The only remaining question for performing this test is the magnitude of c. Obviously, the larger c is, the more central observations are being omitted and the more confident we feel that the two samples are distant from each other. The loss of c observations should lead to loss of power. However, separating the two samples should give us more confidence that the two variances are in fact the same if we do not reject homoskedasticity. This trade off in power was studied by Goldfeld and Quandt using Monte Carlo experiments. Their results recommend the use of $c = 8$ for $n = 30$ and $c = 16$ for $n = 60$. This is a popular test, but assumes that we know how to order the heteroskedasticity. In this case, using X_i. But what if there are more than one regressor on the right hand side? In that case one can order the observations using $\widehat{Y}_i$.

(3) Spearman's Rank Correlation Test: This test ranks the X_i's and the absolute value of the OLS residuals, the e_i's. Then it computes the difference between these rankings, i.e., $d_i = \text{rank}(|e_i|) - \text{rank}(X_i)$. The Spearman-Correlation coefficient is $r = 1 - \left[6 \sum_{i=1}^{n} d_i^2/(n^3 - n)\right]$. Finally, test H_0; the correlation coefficient between the rankings is zero, by computing $t =$

$[r^2(n-2)/(1-r^2)]^{1/2}$ which is t-distributed with $(n-2)$ degrees of freedom. If this t-statistic has a large p-value we do not reject homoskedasticity. Otherwise, we reject homoskedasticity in favor of heteroskedasticity.

(4) Harvey's (1976) Multiplicative Heteroskedasticity Test: If heteroskedasticity is related to X_i, it looks like the Goldfeld and Quandt test or the Spearman rank correlation test would detect it, and the Glejser test would establish its form. In case the form of heteroskedasticity is of the multiplicative type, Harvey (1976) suggests the following test which rewrites (5.17) as

$$\log e_i^2 = \log\sigma^2 + \delta_1 \log Z_{1i} + \ldots + \delta_r \log Z_{ri} + v_i \tag{5.25}$$

where $v_i = \log(e_i^2/\sigma_i^2)$. This disturbance term has an asymptotic distribution that is $\log\chi_1^2$. This random variable has mean -1.2704 and variance 4.9348. Therefore, Harvey suggests performing the regression in (5.25) and testing H_0; $\delta_1 = \delta_2 = \ldots = \delta_r = 0$ by computing the regression sum of squares divided by 4.9348. This statistic is distributed asymptotically as χ_r^2. This is also asymptotically equivalent to an F-test that tests for $\delta_1 = \delta_2 = \ldots = \delta_r = 0$ in the regression given in (5.25). See the F-test described in example 6 of Chapter 4.

(5) Breusch and Pagan (1979) Test: If one knows that $\sigma_i^2 = f(a + b_1 Z_1 + b_2 Z_2 + .. + b_r Z_r)$ but does not know the form of this function f, Breusch and Pagan (1979) suggest the following test for homoskedasticity, i.e., $H_0; b_1 = b_2 = \ldots = b_r = 0$. Compute $\widehat{\sigma}^2 = \sum_{i=1}^n e_i^2/n$, which would be the MLE estimator of σ^2 under homoskedasticity. Run the regression of $e_i^2/\widehat{\sigma}^2$on the Z variables and a constant, and compute half the regression sum of squares. This statistic is distributed as χ_r^2. This is a more general test than the ones discussed earlier in that f does not have to be specified.

(6) White's (1980) Test: Another general test for homoskedasticity where nothing is known about the form of this heteroskedasticity is suggested by White (1980). This test is based on the difference between the variance of the OLS estimates under homoskedasticity and that under heteroskedasticity. For the case of a simple regression with a constant, White shows that this test compares White's $\text{var}(\widehat{\beta}_{OLS})$ given by (5.13) with the usual $\text{var}(\widehat{\beta}_{OLS}) = s^2/\sum_{i=1}^n x_i^2$ under homoskedasticity. This test reduces to running the regression of e_i^2 on a constant, X_i and X_i^2 and computing nR^2. This statistic is distributed as χ_2^2 under the null hypothesis of homoskedasticity. The degrees of freedom correspond to the number of regressors without the constant. If this statistic is not significant, then e_i^2 is not related to X_i and X_i^2 and we can not reject that the variance is constant. Note that if there is no constant in the regression, we run e_i^2 on a constant and X_i^2 only, i.e., X_i is no longer in this regression and the degree of freedom of the test is 1. In general, White's test is based on running e_i^2 on the cross-product of all the X's in the regression being estimated, computing nR^2, and comparing it to the critical value of χ_r^2 where r is the number of regressors in this last regression excluding the constant. For the case of two regressors, X_2 and X_3 and a constant, White's test is again based on nR^2 for the regression of e_i^2 on a constant, $X_2, X_3, X_2^2, X_2X_3, X_3^2$. This statistic is distributed as χ_5^2. White's test is standard using EViews. After running the regression, click on residuals tests then choose White. This software gives the user a choice between including or excluding the cross-product terms like X_2X_3 from the regression. This may be useful when there are many regressors.

A modified Breusch-Pagan test was suggested by Koenker (1981) and Koenker and Bassett (1982). This attempts to improve the power of the Breusch-Pagan test, and make it more robust

to the non-normality of the disturbances. This amounts to multiplying the Breusch-Pagan statistic (half the regression sum of squares) by $2\widehat{\sigma}^4$, and dividing it by the second sample moment of the squared residuals, i.e., $\sum_{i=1}^{n}(e_i^2 - \widehat{\sigma}^2)^2/n$, where $\widehat{\sigma}^2 = \sum_{i=1}^{n} e_i^2/n$. Waldman (1983) showed that if the Z_i's in the Breusch-Pagan test are in fact the X_i's and their cross-products, as in White's test, then this modified Breusch-Pagan test is exactly the nR^2 statistic proposed by White.

White's (1980) test for heteroskedasticity without specifying its form lead to further work on estimators that are more efficient than OLS while recognizing that the efficiency of GLS may not be achievable, see Cragg (1992). Adaptive estimators have been developed by Carroll (1982) and Robinson (1987). These estimators assume no particular form of heteroskedasticity but nevertheless have the same asymptotic distribution as GLS based on the true σ_i^2.

Many Monte Carlo experiments were performed to study the performance of these and other tests of homoskedasticity. One such extensive study is that of Ali and Giaccotto (1984). Six types of heteroskedasticity specifications were considered;

(i) $\sigma_i^2 = \sigma^2$	(ii) $\sigma_i^2 = \sigma^2\|X_i\|$	(iii) $\sigma_i^2 = \sigma^2\|E(Y_i)\|$
(iv) $\sigma_i^2 = \sigma^2 X_i^2$	(v) $\sigma_i^2 = \sigma^2[E(Y_i)]^2$	(vi) $\sigma_i^2 = \sigma^2$ for $i \leq n/2$ and $\sigma_i^2 = 2\sigma^2$ for $i > n/2$

Six data sets were considered, the first three were stationary and the last three were nonstationary (Stationary versus non-stationary regressors, are discussed in Chapter 14). Five models were entertained, starting with a model with one regressor and no intercept and finishing with a model with an intercept and 5 variables. Four types of distributions were imposed on the disturbances. These were normal, t, Cauchy and log normal. The first three are symmetric, but the last one is skewed. Three sample sizes were considered, $n = 10, 25, 40$. Various correlations between the disturbances were also entertained. Among the tests considered were tests 1, 2, 5 and 6 discussed in this section. The results are too numerous to summarize, but some of the major findings are the following: (1) The power of these tests increased with sample size and trendy nature or the variability of the regressors. It also decreased with more regressors and for deviations from the normal distribution. The results were mostly erratic when the errors were autocorrelated. (2) There were ten distributionally robust tests using OLS residuals named TROB which were variants of Glejser's, White's and Bickel's tests. The last one being a non-parametric test not considered in this chapter. These tests were robust to both long-tailed and skewed distributions. (3) None of these tests has any significant power to detect heteroskedasticity which deviates substantially from the true underlying heteroskedasticity. For example, none of these tests was powerful in detecting heteroskedasticity of the sixth kind, i.e., $\sigma_i^2 = \sigma^2$ for $i \leq n/2$ and $\sigma_i^2 = 2\sigma^2$ for $i > n/2$. In fact, the maximum power was 9%. (4) Ali and Giaccotto (1984) recommend any of the TROB tests for practical use. They note that the similarity among these tests is the use of squared residuals rather than the absolute value of the residuals. In fact, they argue that tests of the same form that use absolute value rather than squared residuals are likely to be non-robust and lack power.

Empirical Example: For the Cigarette Consumption Data given in Table 3.2, the OLS regression yields:

$$\begin{array}{lccccc} \log C = & 4.30 & - & 1.34 \ \log P + & 0.17 \ \log Y & \bar{R}^2 = 0.27 \\ & (0.91) & & (0.32) & (0.20) & \end{array}$$

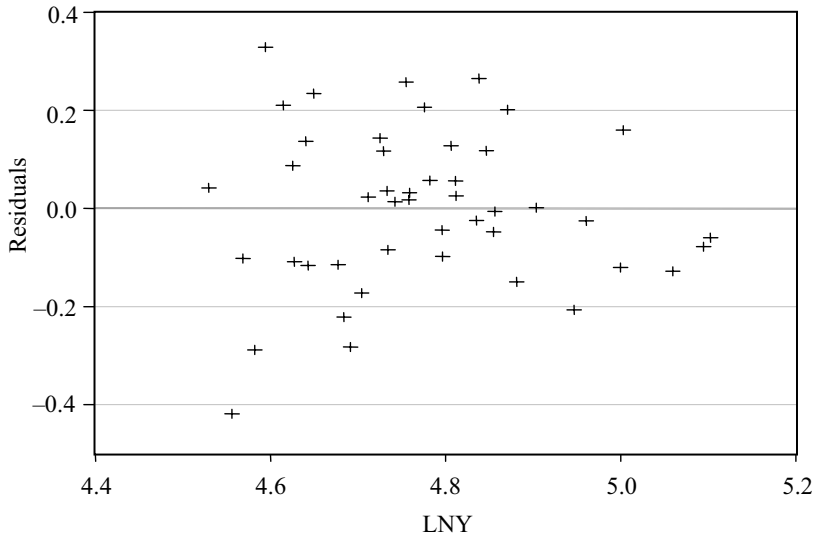


Figure 5.1 Plots of Residuals versus Log Y

Suspecting heteroskedasticity, we plotted the residuals from this regression versus log$Y$ in Figure 5.1. This figure shows the dispersion of the residuals to decrease with increasing log$Y$. Next, we performed several tests for heteroskedasticity studied in this chapter. The first is Glejser's (1969) test. We ran the following regressions:

$$\begin{array}{lrcl} |e_i| = & 1.16 & - & 0.22 \ \log Y \\ & (0.46) & & (0.10) \end{array}$$

$$\begin{array}{lrcl} |e_i| = -0.95 + & 5.13 & & (\log Y)^{-1} \\ & (0.47) & & (2.23) \end{array}$$

$$\begin{array}{lrcl} |e_i| = -2.00 + & 4.65 & & (\log Y)^{-0.5} \\ & (0.93) & & (2.04) \end{array}$$

$$\begin{array}{lrcl} |e_i| = & 2.21 & - & 0.96 \ (\log Y)^{0.5} \\ & (0.93) & & (0.42) \end{array}$$

The t-statistics on the slope coefficient in these regressions are -2.24, 2.30, 2.29 and -2.26, respectively. All are significant with p-values of 0.03, 0.026, 0.027 and 0.029, respectively, indicating the rejection of homoskedasticity.

The second test is the Goldfeld and Quandt (1965) test. The observations are ordered according to log$Y$ and $c = 12$ central observations are omitted. Two regressions are run on the first and last 17 observations. The first regression yields $s_1^2 = 0.04881$ and the second regression yields $s_2^2 = 0.01554$. This is a test of equality of variances and it is based on the ratio of two χ^2 random variables with $17 - 3 = 14$ degrees of freedom. In fact, $s_1^2/s_2^2 = 0.04881/0.01554 = 3.141 \sim F_{14,14}$ under H_0. This has a p-value of 0.02 and rejects H_0 at the 5% level. The third test is the Spearman rank correlation test. First one obtains the rank(logY_i) and rank($|e_i|$) and compute $d_i = \text{rank}|e_i| - \text{rank}|\log Y_i|$. From these $r = 1 - \left[6 \sum_{i=1}^{46} d_i^2/(n^3 - n)\right] = -0.282$ and

Table 5.1 White Heteroskedasticity Test

F-statistic	4.127779	Probability		0.004073
Obs*R-squared	15.65644	Probability		0.007897
Test Equation:				
Dependent Variable:	RESIDˆ2			
Method:	Least Squares			
Sample:	1 46			
Included observations:	46			
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	18.22199	5.374060	3.390730	0.0016
LNP	9.506059	3.302570	2.878382	0.0064
LNPˆ2	1.281141	0.656208	1.952340	0.0579
LNP*LNY	–2.078635	0.727523	–2.857139	0.0068
LNY	–7.893179	2.329386	–3.388523	0.0016
LNYˆ2	0.855726	0.253048	3.381670	0.0016
R-squared	0.340357	Mean dependent var		0.024968
Adjusted R-squared	0.257902	S.D. dependent var		0.034567
S.E. of regression	0.029778	Akaike info criterion		–4.068982
Sum squared resid	0.035469	Schwarz criterion		–3.830464
Log likelihood	99.58660	F-statistic		4.127779
Durbin-Watson stat	1.853360	Prob (F-statistic)		0.004073

$t = [r^2(n-2)/(1-r^2)]^{1/2} = 1.948$. This is distributed as a t with 44 degrees of freedom. This t-statistic has a p-value of 0.058.

The fourth test is Harvey's (1976) multiplicative heteroskedasticity test which is based upon regressing $\log e_i^2$ on $\log(\log Y_i)$

$$\begin{array}{lccc} \log e_i^2 = & 24.85 & - & 19.08 \ \log(\log Y) \\ & (17.25) & & (11.03) \end{array}$$

Harvey's (1976) statistic divides the regression sum of squares which is 14.360 by 4.9348. This yields 2.91 which is asymptotically distributed as χ_1^2 under the null. This has a p-value of 0.088 and does not reject the null of homoskedasticity at the 5% significance level.

The fifth test is the Breusch and Pagan (1979) test which is based on the regression of $e_i^2/\widehat{\sigma}^2$ (where $\widehat{\sigma}^2 = \sum_{i=1}^{46} e_i^2/46 = 0.024968$) on $\log Y_i$. The test-statistic is half the regression sum of squares $= (10.971 \div 2) = 5.485$. This is distributed as χ_1^2 under the null hypothesis. This has a p-value of 0.019 and rejects the null of homoskedasticity.

Finally, White's (1980) test for heteroskedasticity is performed which is based on the regression of e_i^2 on $\log P$, $\log Y$, $(\log P)^2$, $(\log Y)^2$, $(\log P)(\log Y)$ and a constant. This is shown in Table 5.1 using EViews. The test-statistic is $nR^2 = (46)(0.3404) = 15.66$ which is distributed as χ_5^2. This has a p-value of 0.008 and rejects the null of homoskedasticity. Except for Harvey's test, all the tests performed indicate the presence of heteroskedasticity. This is true despite the fact that the data are in logs, and both consumption and income are expressed in per capita terms.

Table 5.2 White Heteroskedasticity-Consistent Standard Errors

Dependent Variable:	LNC
Method:	Least Squares
Sample:	1 46
Included observations:	46

White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.299662	1.095226	3.925821	0.0003
LNP	–1.338335	0.343368	–3.897671	0.0003
LNY	0.172386	0.236610	0.728565	0.4702

R-squared	0.303714	Mean dependent var	4.847844
Adjusted R-squared	0.271328	S.D. dependent var	0.191458
S.E. of regression	0.163433	Akaike info criterion	–0.721834
Sum squared resid	1.148545	Schwarz criterion	–0.602575
Log likelihood	19.60218	F-statistic	9.378101
Durbin-Watson stat	2.315716	Prob (F-statistic)	0.000417

White's heteroskedasticity-consistent estimates of the variances are as follows:

$$\begin{array}{lccccc} \log C = & 4.30 & - & 1.34 \ \log P + & 0.17 \ \log Y \\ & (1.10) & & (0.34) & & (0.24) \end{array}$$

These are given in Table 5.2 using EViews. Note that in this case all of the heteroskedasticity-consistent standard errors are larger than those reported using a standard OLS package, but this is not necessarily true for other data sets.

In section 5.4, we described the Jarque and Bera (1987) test for normality of the disturbances. For this cigarette consumption regression, Figure 5.2 gives the histogram of the residuals along with descriptive statistics of these residuals including their mean, median, skewness and kurtosis.

This is done using EViews. The measure of skewness S is estimated to be -0.184 and the measure of kurtosis κ is estimated to be 2.875 yielding a Jarque-Bera statistic of

$$JB = 46\left[\frac{(-0.184)^2}{6} + \frac{(2.875-3)^2}{24}\right] = 0.29.$$

This is distributed as χ^2_2 under the null hypothesis of normality and has a p-value of 0.865. Hence we do not reject that the distribution of the disturbances is symmetric and has a kurtosis of 3.

5.6 Autocorrelation

Violation of assumption 3 means that the disturbances are correlated, i.e., $E(u_i u_j) = \sigma_{ij} \neq 0$, for $i \neq j$, and $i, j = 1, 2, \ldots, n$. Since u_i has zero mean, $E(u_i u_j) = \text{cov}(u_i, u_j)$ and this is denoted by σ_{ij}. This correlation is more likely to occur in time-series than in cross-section studies. Consider estimating the consumption function of a random sample of households. An unexpected event, like a visit of family members will increase the consumption of this household. However, this

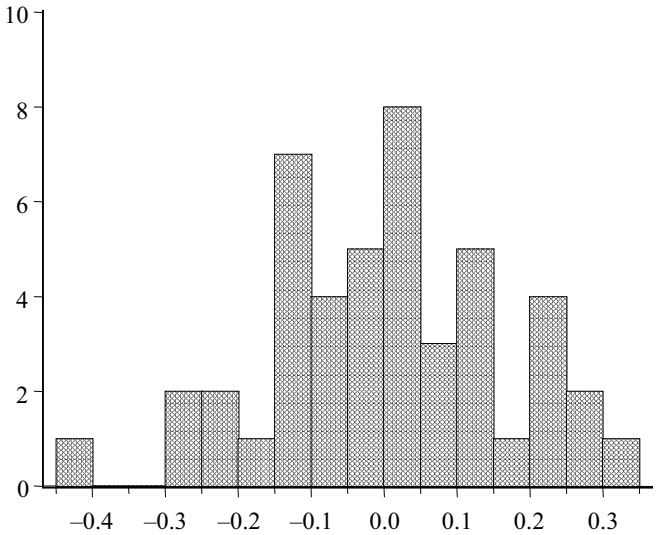


Figure 5.2 Normality Test (Jarque-Bera)

positive disturbance need not be correlated to the disturbances affecting consumption of other randomly drawn households. However, if we were estimating this consumption function using aggregate time-series data for the U.S., then it is very likely that a recession year affecting consumption negatively this year may have a carry over effect to the next few years. A shock to the economy like an oil embargo in 1973 is likely to affect the economy for several years. A labor strike this year may affect production for the next few years. Therefore, we will switch the i and j subscripts to t and s denoting time-series observations $t, s = 1, 2, \ldots, T$ and the sample size will be denoted by T rather than n. This covariance term is symmetric, so that $\sigma_{12} = E(u_1 u_2) = E(u_2 u_1) = \sigma_{21}$. Hence, only $T(T-1)/2$ distinct σ_{ts}'s have to be estimated. For example, if $T = 3$, then σ_{12}, σ_{13} and σ_{23} are the distinct covariance terms. However, it is hopeless to estimate $T(T-1)/2$ covariances (σ_{ts}) with only T observations. Therefore, more structure on these σ_{ts}'s need to be imposed. A popular assumption is that the u_t's follow a first-order autoregressive process denoted by AR(1):

$$u_t = \rho u_{t-1} + \epsilon_t \quad t = 1, 2, \ldots, T \tag{5.26}$$

where ϵ_t is IID$(0, \sigma_\epsilon^2)$. It is autoregressive because u_t is related to its lagged value u_{t-1}. One can also write (5.26), for period $t-1$, as

$$u_{t-1} = \rho u_{t-2} + \epsilon_{t-1} \tag{5.27}$$

and substitute (5.27) in (5.26) to get

$$u_t = \rho^2 u_{t-2} + \rho \epsilon_{t-1} + \epsilon_t \tag{5.28}$$

Note that the power of ρ and the subscript of u or ϵ always sum to t. By continuous substitution of this form, one ultimately gets

$$u_t = \rho^t u_0 + \rho^{t-1} \epsilon_1 + .. + \rho \epsilon_{t-1} + \epsilon_t \tag{5.29}$$

This means that u_t is a function of current and past values of ϵ_t and u_0 where u_0 is the initial value of u_t. If u_0 has zero mean, then u_t has zero mean. This follows from (5.29) by taking expectations. Also, from (5.26)

$$\text{var}(u_t) = \rho^2 \text{var}(u_{t-1}) + \text{var}(\epsilon_t) + 2\rho \text{cov}(u_{t-1}, \epsilon_t) \tag{5.30}$$

Using (5.29), u_{t-1} is a function of ϵ_{t-1}, past values of ϵ_{t-1} and u_0. Since u_0 is independent of the ϵ's, and the ϵ's are themselves not serially correlated, then u_{t-1} is independent of ϵ_t. This means that $\text{cov}(u_{t-1}, \epsilon_t) = 0$. Furthermore, for u_t to be homoskedastic, $\text{var}(u_t) = \text{var}(u_{t-1}) = \sigma_u^2$, and (5.30) reduces to $\sigma_u^2 = \rho^2 \sigma_u^2 + \sigma_\epsilon^2$, which when solved for σ_u^2 gives:

$$\sigma_u^2 = \sigma_\epsilon^2/(1-\rho^2) \tag{5.31}$$

Hence, $u_0 \sim (0, \sigma_\epsilon^2/(1-\rho^2))$ for the u's to have zero mean and homoskedastic disturbances. Multiplying (5.26) by u_{t-1} and taking expected values, one gets

$$E(u_t u_{t-1}) = \rho E(u_{t-1}^2) + E(u_{t-1}\epsilon_t) = \rho \sigma_u^2 \tag{5.32}$$

since $E(u_{t-1}^2) = \sigma_u^2$ and $E(u_{t-1}\epsilon_t) = 0$. Therefore, $\text{cov}(u_t, u_{t-1}) = \rho\sigma_u^2$, and the correlation coefficient between u_t and u_{t-1} is $\text{correl}(u_t, u_{t-1}) = \text{cov}(u_t, u_{t-1})/\sqrt{\text{var}(u_t)\text{var}(u_{t-1})} = \rho\sigma_u^2/\sigma_u^2 = \rho$. Since ρ is a correlation coefficient, this means that $-1 \leq \rho \leq 1$. In general, one can show that

$$\text{cov}(u_t, u_s) = \rho^{|t-s|}\sigma_u^2 \quad t, s = 1, 2, \ldots, T \tag{5.33}$$

see problem 6. This means that the correlation between u_t and u_{t-r} is ρ^r, which is a fraction raised to an integer power, i.e., the correlation is decaying between the disturbances the further apart they are. This is reasonable in economics and may be the reason why this autoregressive form (5.26) is so popular. One should note that this is not the only form that would correlate the disturbances across time. In Chapter 14, we will consider other forms like the Moving Average (MA) process, and higher order Autoregressive Moving Average (ARMA) processes, but these are beyond the scope of this chapter.

Consequences for OLS

How is the OLS estimator affected by the violation of the no autocorrelation assumption among the disturbances? The OLS estimator is still unbiased and consistent since these properties rely on assumptions 1 and 4 and have nothing to do with assumption 3. For the simple linear regression, using (5.2), the variance of $\widehat{\beta}_{OLS}$ is now

$$\begin{aligned} \text{var}(\widehat{\beta}_{OLS}) &= \text{var}(\textstyle\sum_{t=1}^T w_t u_t) = \sum_{t=1}^T \sum_{s=1}^T w_t w_s \text{cov}(u_t, u_s) \\ &= \sigma_u^2 / \textstyle\sum_{t=1}^T x_t^2 + \sum\sum_{t \neq s} w_t w_s \rho^{|t-s|} \sigma_u^2 \end{aligned} \tag{5.34}$$

where $\text{cov}(u_t, u_s) = \rho^{|t-s|}\sigma_u^2$ as explained in (5.33). Note that the first term in (5.34) is the usual variance of $\widehat{\beta}_{OLS}$ under the classical case. The second term in (5.34) arises because of the correlation between the u_t's. Hence, the variance of OLS computed from a regression package, i.e., $s^2/\sum_{t=1}^T x_t^2$ is a wrong estimate of the variance of $\widehat{\beta}_{OLS}$ for two reasons. First, it is using the wrong formula for the variance, i.e., $\sigma_u^2/\sum_{t=1}^T x_t^2$ rather than (5.34). The latter depends on ρ through the extra term in (5.34). Second, one can show, see problem 7, that $E(s^2) \neq \sigma_u^2$ and will

involve ρ as well as σ_u^2. Hence, s^2 is not unbiased for σ_u^2 and $s^2/\sum_{t=1}^{T} x_t^2$ is a biased estimate of $\text{var}(\widehat{\beta}_{OLS})$. The direction and magnitude of this bias depends on ρ and the regressor. In fact, if ρ is positive, and the x_t's are themselves positively autocorrelated, then $s^2/\sum_{t=1}^{T} x_t^2$ understates the true variance of $\widehat{\beta}_{OLS}$. This means that the confidence interval for β is tighter than it should be and the t-statistic for H_0; $\beta = 0$ is overblown, see problem 8. As in the heteroskedastic case, but for completely different reasons, any inference based on $\text{var}(\widehat{\beta}_{OLS})$ reported from the standard regression packages will be misleading if the u_t's are serially correlated.

Newey and West (1987) suggested a simple heteroskedasticity and autocorrelation-consistent covariance matrix for the OLS estimator without specifying the functional form of the serial correlation. The basic idea extends White's (1980) replacement of heteroskedastic variances with squared OLS residuals e_t^2 by additionally including products of least squares residuals $e_t e_{t-s}$ for $s = 0, \pm 1, \ldots, \pm p$ where p is the maximum order of serial correlation we are willing to assume. The consistency of this procedure relies on p being very small relative to the number of observations T. This is consistent with popular serial correlation specifications considered in this chapter where the autocorrelation dies out quickly as j increases. Newey and West (1987) allow the higher order covariance terms to receive diminishing weights. This Newey-West option for the least squares estimator is available using EViews. Andrews (1991) warns about the unreliability of such standard error corrections in some circumstances. Wooldridge (1991) shows that it is possible to construct serially correlated robust F-statistics for testing joint hypotheses as considered in Chapter 4. However, these are beyond the scope of this book.

Is OLS still BLUE? In order to determine the BLU estimator in this case, we lag the regression equation once, multiply it by ρ, and subtract it from the original regression equation, we get

$$Y_t - \rho Y_{t-1} = \alpha(1-\rho) + \beta(X_t - \rho X_{t-1}) + \epsilon_t \quad t = 2, 3, \ldots, T \tag{5.35}$$

This transformation, known as the Cochrane-Orcutt (1949) transformation, reduces the disturbances to classical errors. Therefore, OLS on the resulting regression renders the estimates BLU, i.e., run $\widetilde{Y}_t = Y_t - \rho Y_{t-1}$ on a constant and $\widetilde{X}_t = X_t - \rho X_{t-1}$, for $t = 2, 3, \ldots, T$. Note that we have lost one observation by lagging, and the resulting estimators are BLUE only for linear combinations of $(T-1)$ observations in Y.[1] Prais and Winsten (1954) derive the BLU estimators for linear combinations of T observations in Y. This entails recapturing the initial observation as follows: (i) Multiply the first observation of the regression equation by $\sqrt{1-\rho^2}$;

$$\sqrt{1-\rho^2}Y_1 = \alpha\sqrt{1-\rho^2} + \beta\sqrt{1-\rho^2}X_1 + \sqrt{1-\rho^2}u_1$$

(ii) add this transformed initial observation to the Cochrane-Orcutt transformed observations for $t = 2, \ldots, T$ and run the regression on the T observations rather than the $(T-1)$ observations. See Chapter 9, for a formal proof of this result. Note that

$$\widetilde{Y}_1 = \sqrt{1-\rho^2}Y_1$$

and

$$\widetilde{Y}_t = Y_t - \rho Y_{t-1} \quad \text{for } t = 2, \ldots, T$$

Similarly, $\widetilde{X}_1 = \sqrt{1-\rho^2}X_1$ and $\widetilde{X}_t = X_t - \rho X_{t-1}$ for $t = 2, \ldots, T$. The constant variable $C_t = 1$ for $t = 1, \ldots, T$ is now a new variable $\widetilde{C}_t$ which takes the values $\widetilde{C}_1 = \sqrt{1-\rho^2}$ and $\widetilde{C}_t = (1-\rho)$ for $t = 2, \ldots, T$. Hence, the Prais-Winsten procedure is the regression of $\widetilde{Y}_t$ on $\widetilde{C}_t$ and $\widetilde{X}_t$

without a constant. It is obvious that the resulting BLU estimators will involve ρ and are therefore, different from the usual OLS estimators except in the case where $\rho = 0$. Hence, OLS is no longer BLUE. Furthermore, we need to know ρ in order to obtain the BLU estimators. In applied work, ρ is not known and has to be estimated, in which case the Prais-Winsten regression is no longer BLUE since it is based on an estimate of ρ rather than the true ρ itself. However, as long as $\widehat{\rho}$ is a consistent estimate for ρ then this is a sufficient condition for the corresponding estimates of α and β in the next step to be asymptotically efficient, see Chapter 9. We now turn to various methods of estimating ρ.

(1) The Cochrane-Orcutt (1949) Method: This method starts with an initial estimate of ρ, the most convenient is 0, and minimizes the residual sum of squares in (5.35). This gives us the OLS estimates of α and β. Then we substitute $\widehat{\alpha}_{OLS}$ and $\widehat{\beta}_{OLS}$ in (5.35) and we get

$$e_t = \rho e_{t-1} + \epsilon_t \quad t = 2, \ldots, T \tag{5.36}$$

where e_t denotes the OLS residual. An estimate of ρ can be obtained by minimizing the residual sum of squares in (5.36) or running the regression of e_t on e_{t-1} without a constant. The resulting estimate of ρ is $\widehat{\rho}_{co} = \sum_{t=2}^{T} e_t e_{t-1} / \sum_{t=2}^{T} e_{t-1}^2$ where both summations run over $t = 2, 3, \ldots, T$. The second step of the Cochrane-Orcutt procedure (2SCO) is to perform the regression in (5.35) with $\widehat{\rho}_{co}$ instead of ρ. One can iterate this procedure (ITCO) by computing new residuals based on the new estimates of α and β and hence a new estimate of ρ from (5.36), and so on, until convergence. Both the 2SCO and the ITCO are asymptotically efficient, the argument for iterating must be justified in terms of small sample gains.

(2) The Hilderth-Lu (1960) Search Procedure: ρ is between -1 and 1. Therefore, this procedure searches over this range, i.e., using values of ρ say between -0.9 and 0.9 in intervals of 0.1. For each ρ, one computes the regression in (5.35) and reports the residual sum of squares corresponding to that ρ. The minimum residual sum of squares gives us our choice of ρ and the corresponding regression gives us the estimates of α, β and σ^2. One can refine this procedure around the best ρ found in the first stage of the search. For example, suppose that $\rho = 0.6$ gave the minimum residual sum of squares, one can search next between 0.51 and 0.69 in intervals of 0.01. This search procedure guards against a local minimum. Since the likelihood in this case contains ρ as well as σ^2 and α and β, this search procedure can be modified to maximize the likelihood rather than minimize the residual sum of squares, since the two criteria are no longer equivalent. The maximum value of the likelihood will give our choice of ρ and the corresponding estimates of α, β and σ^2.

(3) Durbin's (1960) Method: One can rearrange (5.35) by moving Y_{t-1} to the right hand side, i.e.,

$$Y_t = \rho Y_{t-1} + \alpha(1 - \rho) + \beta X_t - \rho\beta X_{t-1} + \epsilon_t \tag{5.37}$$

and running OLS on (5.37). The error in (5.37) is classical, and the presence of Y_{t-1} on the right hand side reminds us of the contemporaneously uncorrelated case discussed under the violation of assumption 4. For that violation, we have shown that unbiasedness is lost, but not consistency. Hence, the estimate of ρ as a coefficient of Y_{t-1} is biased but consistent. This is the Durbin estimate of ρ, call it $\widehat{\rho}_D$. Next, the second step of the Cochrane-Orcutt procedure is performed using this estimate of ρ.

(4) Beach-MacKinnon (1978) Maximum Likelihood Procedure: Beach and MacKinnon (1978) derived a cubic equation in ρ which maximizes the likelihood function concentrated with respect to α, β, and σ^2. With this estimate of ρ, denoted by $\widehat{\rho}_{BM}$, one performs the Prais-Winsten procedure in the next step.

Correcting for serial correlation is not without its critics. Mizon (1995) argues this point forcefully in his article entitled "A simple message for autocorrelation correctors: Don't." The main point being that serial correlation is a symptom of dynamic misspecification which is better represented using a general unrestricted dynamic specification.

Monte Carlo Results

Rao and Griliches (1969) performed a Monte Carlo study using an autoregressive X_t, and various values of ρ. They found that OLS is still a viable estimator as long as $|\rho| < 0.3$, but if $|\rho| > 0.3$, then it pays to perform procedures that correct for serial correlation based on an estimator of ρ. Their recommendation was to compute a Durbin's estimate of ρ in the first step and to do the Prais-Winsten procedure in the second step. Maeshiro (1976, 1979) found that if the X_t series is trended, which is usual with economic data, then OLS outperforms 2SCO, but not the two-step Prais-Winsten (2SPW) procedure that recaptures the initial observation. These results were confirmed by Park and Mitchell (1980) who performed an extensive Monte Carlo using trended and untrended X_t's. Their basic findings include the following: (i) For trended X_t's, OLS beats 2SCO, ITCO and even a Cochrane-Orcutt procedure that is based on the true ρ. However, OLS was beaten by 2SPW, iterative Prais-Winsten (ITPW), and Beach-MacKinnon (BM). Their conclusion is that one should not use regressions based on $(T-1)$ observations as in Cochrane and Orcutt. (ii) Their results find that the ITPW procedure is the recommended estimator beating 2SPW and BM for high values of true ρ, for both trended as well as nontrended X_t's. (iii) Test of hypotheses regarding the regression coefficients performed miserably for all estimators based on an estimator of ρ. The results indicated less bias in standard error estimation for ITPW, BM and 2SPW than OLS. However, the tests based on these standard errors still led to a high probability of type I error for all estimation procedures.

Testing for Autocorrelation

So far, we have studied the properties of OLS under the violation of assumption 3. We have derived asymptotically efficient estimators of the coefficients based on consistent estimators of ρ and studied their small sample properties using Monte Carlo experiments. Next, we focus on the problem of detecting this autocorrelation between the disturbances. A popular diagnostic for detecting such autocorrelation is the Durbin and Watson (1951) statistic[2]

$$d = \sum_{t=2}^{T}(e_t - e_{t-1})^2 / \sum_{t=1}^{T} e_t^2 \tag{5.38}$$

If this was based on the true u_t's and T was very large then d can be shown to tend in the limit as T gets large to $2(1-\rho)$, see problem 9. This means that if $\rho \to 0$, then $d \to 2$; if $\rho \to 1$, then $d \to 0$ and if $\rho \to -1$, then $d \to 4$. Therefore, a test for H_0; $\rho = 0$, can be based on whether d is close to 2 or not. Unfortunately, the critical values of d depend upon the X_t's, and these vary from one data set to another. To get around this, Durbin and Watson established upper (d_U) and lower (d_L) bounds for this critical value. Figure 5.3 shows these bounds.

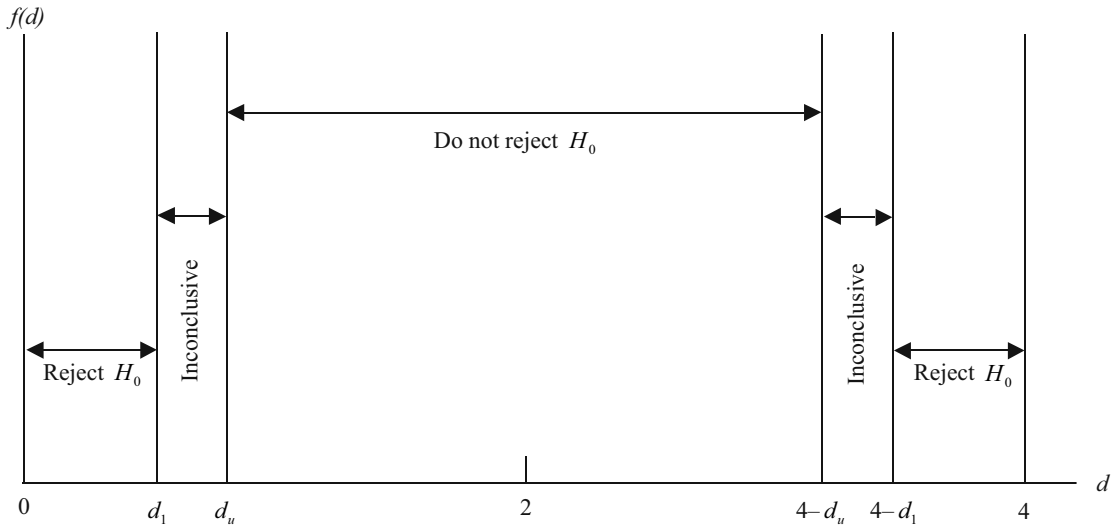


Figure 5.3 Durbin-Watson Critical Values

It is obvious that if the observed d is less than d_L, or larger than $4 - d_L$, we reject H_0. If the observed d is between d_U and $4 - d_U$, then we do not reject H_0. If d lies in any of the two indeterminant regions, then one should compute the exact critical values depending on X_t. Most regression packages report the Durbin-Watson statistic. SHAZAM gives the exact p-value for this d-statistic. If one is interested in a single sided test, say H_0; $\rho = 0$ versus $H_1; \rho > 0$ then one would reject H_0 if $d < d_L$, and not reject H_0 if $d > d_U$. If $d_L < d < d_U$, then the test is inconclusive. Similarly for testing $H_0; \rho = 0$ versus $H_1; \rho < 0$, one computes $(4 - d)$ and follow the steps for testing against positive autocorrelation. Durbin and Watson tables for d_L and d_U covered samples sizes from 15 to 100 and a maximum of 5 regressors. Savin and White (1977) extended these tables for $6 \leq T \leq 200$ and up to 10 regressors.

The Durbin-Watson statistic has several limitations. We discussed the inconclusive region and the computation of exact critical values. The Durbin-Watson statistic is appropriate when there is a constant in the regression. In case there is no constant in the regression, see Farebrother (1980). Also, the Durbin-Watson statistic is inappropriate when there are lagged values of the dependent variable among the regressors. We now turn to an alternative test for serial correlation that does not have these limitations and that is also easy to apply. This test was derived by Breusch (1978) and Godfrey (1978) and is known as the Breusch-Godfrey test for zero first-order serial correlation. This is a Lagrange Multiplier test that amounts to running the regression of the OLS residuals e_t on e_{t-1} and the original regressors in the model. The test statistic is TR^2. Its distribution under the null is χ^2_1. In this case, the regressors are a constant and X_t, and the test checks whether the coefficient of e_{t-1} is significant. The beauty of this test is that (i) it is the same test for first-order serial correlation, whether the disturbances are Moving Average of order one MA(1) or AR(1). (ii) This test is easily generalizable to higher autoregressive or Moving Average schemes. For second-order serial correlation, like MA(2) or AR(2) one includes two lags of the residuals on the right hand side; i.e., both e_{t-1} and e_{t-2}. (iii) This test is still valid even when lagged values of the dependent variable are present among the regressors, see Chapter 6. The Breusch and Godfrey test is standard using EViews and it prompts the user

with a choice of the number of lags of the residuals to include among the regressors to test for serial correlation. You click on residuals, then tests and choose Breusch-Godfrey. Next, you input the number of lagged residuals you want to include.

In conclusion, we focus on first differencing the data as a possible solution for getting rid of serial correlation in the errors. Some economic behavioral equations have variables in first difference form, but other equations are first differenced for estimation purposes. In the latter case, if the original disturbances were not autocorrelated, (or even correlated, with $\rho \neq 1$), then the transformed disturbances are serially correlated. After all, first differencing the disturbances is equivalent to setting $\rho = 1$ in $u_t - \rho u_{t-1}$, and this new disturbance $u_t^* = u_t - u_{t-1}$ has u_{t-1} in common with $u_{t-1}^* = u_{t-1} - u_{t-2}$, making $E(u_t^* u_{t-1}^*) = -E(u_{t-1}^2) = -\sigma_u^2$. However, one could argue that if ρ is large and positive, first differencing the data may not be a bad solution. Rao and Miller (1971) calculated the variance of the BLU estimator correcting for serial correlation, for various guesses of ρ. They assume a true ρ of 0.2, and an autoregressive X_t

$$X_t = \lambda X_{t-1} + w_t \quad \text{with } \lambda = 0, 0.4, 0.8. \tag{5.39}$$

They find that OLS (or a guess of $\rho = 0$), performs better than first differencing the data, and is pretty close in terms of efficiency to the true BLU estimator for trended X_t ($\lambda = 0.8$). However, the performance of OLS deteriorates as λ declines to 0.4 and 0, with respect to the true BLU estimator. This supports the Monte Carlo finding by Rao and Griliches that for $|\rho| < 0.3$, OLS performs reasonably well relative to estimators that correct for serial correlation. However, the first-difference estimator, i.e., a guess of $\rho = 1$, performs badly for trended X_t ($\lambda = 0.8$) giving the worst efficiency when compared to any other guess of ρ. Only when the X_t's are less trended ($\lambda = 0.4$) or random ($\lambda = 0$), does the efficiency of the first-difference estimator improve. However, even for those cases one can do better by guessing ρ. For example, for $\lambda = 0$, one can always do better than first differencing by guessing any positive ρ less than 1. Similarly, for true $\rho = 0.6$, a higher degree of serial correlation, Rao and Miller (1971) show that the performance of OLS deteriorates, while that of the first difference improves. However, one can still do better than first differencing by guessing in the interval $(0.4, 0.9)$. This gain in efficiency increases with trended X_t's.

Empirical Example: Table 5.3 gives the U.S. Real Personal Consumption Expenditures (C) and Real Disposable Personal Income (Y) from the Economic Report of the President over the period 1950-1993. This data set is available as CONSUMP.DAT on the Springer web site.

The OLS regression yields:

$$\begin{aligned} C_t = &-65.80 + \ 0.916 \ Y_t + \text{residuals} \\ &(90.99) \quad (0.009) \end{aligned}$$

Figure 5.4 plots the actual, fitted and residuals using EViews. This shows positive serial correlation with a string of positive residuals followed by a string of negative residuals followed by positive residuals. The Durbin-Watson statistic is $d = 0.461$ which is much smaller than the lower bound $d_L = 1.468$ for $T = 44$ and one regressor. Therefore, we reject the null hypothesis of $H_0; \rho = 0$ at the 5% significance level.

Regressing OLS residuals on their lagged values yields

$$\begin{aligned} e_t = \ &0.792 \ e_{t-1} + \text{residuals} \\ &(0.106) \end{aligned}$$

Table 5.3 U.S. Consumption Data, 1950–1993

C = Real Personal Consumption Expenditures (in 1987 dollars)
Y = Real Disposable Personal Income (in 1987 dollars)

YEAR	Y	C	YEAR	Y	C
1950	6284	5820	1972	10414	9425
1951	6390	5843	1973	11013	9752
1952	6476	5917	1974	10832	9602
1953	6640	6054	1975	10906	9711
1954	6628	6099	1976	11192	10121
1955	6879	6325	1977	11406	10425
1956	7080	6440	1978	11851	10744
1957	7114	6465	1979	12039	10876
1958	7113	6449	1980	12005	10746
1959	7256	6658	1981	12156	10770
1960	7264	6698	1982	12146	10782
1961	7382	6740	1983	12349	11179
1962	7583	6931	1984	13029	11617
1963	7718	7089	1985	13258	12015
1964	8140	7384	1986	13552	12336
1965	8508	7703	1987	13545	12568
1966	8822	8005	1988	13890	12903
1967	9114	8163	1989	14005	13029
1968	9399	8506	1990	14101	13093
1969	9606	8737	1991	14003	12899
1970	9875	8842	1992	14279	13110
1971	10111	9022	1993	14341	13391

Source: Economic Report of the President

The second step of the Cochrane-Orcutt (1949) procedure based on $\widehat{\rho} = 0.792$ yields the following regression:

$$\begin{aligned}(C_t - 0.792C_{t-1}) = 168.49 + \ & 0.926 \ (Y_t - 0.792Y_{t-1}) + \text{residuals} \\ & (301.7) \quad (0.027)\end{aligned}$$

The two-step Prais-Winsten (1954) procedure yields $\widetilde{\rho}_{PW} = 0.707$ with standard error (0.110) and the regression estimates are given by

$$\begin{aligned}C_t = -52.63 + \ & 0.917 \ Y_t + \ \text{residuals} \\ & (181.9) \quad (0.017)\end{aligned}$$

Iterative Prais-Winsten yields

$$\begin{aligned}C_t = -48.40 + \ & 0.916 \ Y_t + \text{residuals} \\ & (191.1) \quad (0.018)\end{aligned}$$

The Maximum-Likelihood procedure yields $\widehat{\rho}_{MLE} = 0.788$ with standard error (0.11) and the resulting estimates are given by

$$\begin{aligned}C_t = -24.82 + \ & 0.915 \ Y_t + \text{residuals} \\ & (233.5) \quad (0.022)\end{aligned}$$

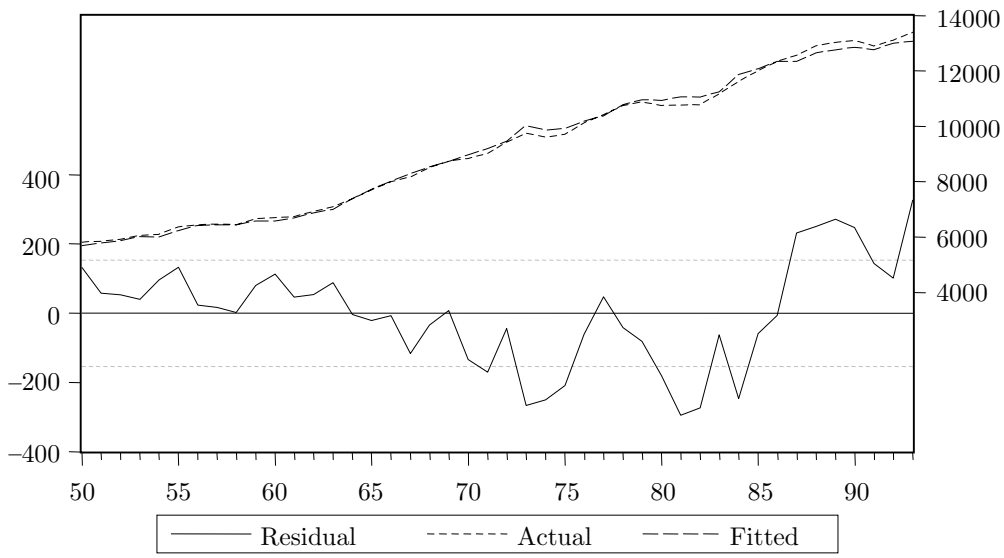


Figure 5.4 Consumption and Disposable Income

Table 5.4 Breusch-Godfrey LM Test

F-statistic	53.45697	Probability		0.00000
Obs*R-squared	24.90136	Probability		0.000001
Test Equation:				
Dependent Variable:	RESID			
Method:	Least Squares			
Presample missing value lagged residuals set to zero				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	–31.12035	60.82348	–0.511650	0.6116
Y	0.003641	0.005788	0.629022	0.5328
RESID(–1)	0.800205	0.109446	7.311428	0.0000
R-squared	0.565940	Mean dependent var		9.83E-13
Adjusted R-squared	0.544766	S.D. dependent var		151.8049
S.E. of regression	102.4243	Akaike info criterion		12.16187
Sum squared resid	430120.2	Schwarz criterion		12.28352
Log likelihood	–264.5612	F-statistic		26.72849
Durbin-Watson stat	1.973552	Prob (F-statistic)		0.000000

Table 5.5 Newey-West Standard Errors

Dependent Variable:	CONSUM
Method:	Least Squares
Sample:	1950 1993
Included observations:	44

Newey-West HAC Standard Errors & Covariance (lag truncation=3)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	–65.79582	133.3454	–0.493424	0.6243
Y	0.915623	0.015458	59.23190	0.0000

R-squared	0.996267	Mean dependent var	9250.545
Adjusted R-squared	0.996178	S.D. dependent var	2484.624
S.E. of regression	153.6015	Akaike info criterion	12.95099
Sum squared resid	990923.1	Schwarz criterion	13.03209
Log likelihood	–282.9218	F-statistic	11209.21
Durbin-Watson stat	0.460778	Prob (F-statistic)	0.000000

Durbin's (1960) Method yields the following regression

$$C_t = \underset{(0.10)}{0.80}\ C_{t-1} - \underset{(59.8)}{40.79} + \underset{(0.09)}{0.72}\ Y_t - \underset{(0.13)}{0.53}\ Y_{t-1} + \text{residuals}$$

Therefore, Durbin's estimate of ρ is given by $\widehat{\rho}_D = 0.80$.

The Breusch (1978) and Godfrey (1978) regression that tests for first-order serial correlation is given in Table 5.4.

This yields

$$e_t = \underset{(60.82)}{-31.12} + \underset{(0.006)}{0.004}\ Y_t + \underset{(0.109)}{0.800}\ e_{t-1} + \text{residuals}$$

The test statistic is TR^2 which yields $43 \times (0.565) = 24.9$. This is distributed as χ^2_1 under H_0; $\rho = 0$. This rejects the null hypothesis of no serial correlation with a p-value of 0.000001 shown in Table 5.4. The Newey-West heteroskedasticity and autocorrelation-consistent standard errors for least squares with a three-year lag truncation are given by

$$C_t = \underset{(133.3)}{-65.80} + \underset{(0.015)}{0.916}\ Y_t + \text{residuals}$$

This is given in Table 5.5 using EViews. Note that both standard errors are now larger than those reported by least squares. But once again, this is not necessarily the case for other data sets.

Notes

1. A computational warning is in order when one is applying the Cochrane-Orcutt transformation to cross-section data. Time-series data has a natural ordering which is generally lacking in cross-section data. Therefore, one should be careful in applying the Cochrane-Orcutt transformation to cross-section data since it is not invariant to the ordering of the observations.

2. Another test for serial correlation can be obtained as a by-product of maximum likelihood estimation. The maximum likelihood estimator of ρ has a normal limiting distribution with mean ρ and variance $(1-\rho^2)/T$. Hence, one can compute $\widehat{\rho}_{MLE}/[(1-\widehat{\rho}^2_{MLE})/T]^{1/2}$and compare it to critical values from the normal distribution.

Problems

1. For the simple linear regression with heteroskedasticity, i.e., $E(u_i^2)=\sigma_i^2$, show that $E(s^2)$ is a function of the σ_i^2's?

2. For the simple linear regression with heteroskedasticity of the form $E(u_i^2)=\sigma_i^2=bx_i^2$ where $b>0$, show that $E(s^2/\sum_{i=1}^n x_i^2)$ understates the variance of $\widehat{\beta}_{OLS}$ which is

$$\sum_{i=1}^n x_i^2\sigma_i^2/(\sum_{i=1}^n x_i^2)^2.$$

3. *Weighted Least Squares.* This is based on Kmenta (1986).

   (a) Solve the two equations in (5.11) and show that the solution is given by (5.12).

   (b) Show that

$$\begin{aligned}\text{var}(\widetilde{\beta}) &= \frac{\sum_{i=1}^n(1/\sigma_i^2)}{[\sum_{i=1}^n X_i^2/\sigma_i^2][\sum_{i=1}^n(1/\sigma_i^2)]-[\sum_{i=1}^n(X_i/\sigma_i^2)]^2}\\ &= \frac{\sum_{i=1}^n w_i^*}{(\sum_{i=1}^n w_i^*X_i^2)(\sum_{i=1}^n w_i^*)-(\sum_{i=1}^n w_i^*X_i)^2}\\ &= \frac{1}{\sum_{i=1}^n w_i^*(X_i-\bar{X}^*)^2}\end{aligned}$$

   where $w_i^*=(1/\sigma_i^2)$ and $\bar{X}^*=\sum_{i=1}^n w_i^*X_i/\sum_{i=1}^n w_i^*$.

4. *Relative Efficiency of OLS Under Heteroskedasticity.* Consider the simple linear regression with heteroskedasticity of the form $\sigma_i^2=\sigma^2X_i^\delta$ where $X_i=1,2,\ldots,10$.

   (a) Compute $\text{var}(\widehat{\beta}_{OLS})$ for $\delta=0.5,1,1.5$ and 2.

   (b) Compute $\text{var}(\widetilde{\beta}_{BLUE})$ for $\delta=0.5,1,1.5$ and 2.

   (c) Compute the efficiency of $\widehat{\beta}_{OLS}=\text{var}(\widetilde{\beta}_{BLUE})/\text{var}(\widehat{\beta}_{OLS})$ for $\delta=0.5,1,1.5$ and 2. What happens to this efficiency measure as δ increases?

5. Consider the simple regression with only a constant $y_i=\alpha+u_i$ for $i=1,2,\ldots,n$; where the u_i's are independent with mean zero and $\text{var}(u_i)=\sigma_1^2$ for $i=1,2,\ldots,n_1$; and $\text{var}(u_i)=\sigma_2^2$ for $i=n_1+1,\ldots,n_1+n_2$ with $n=n_1+n_2$.

   (a) Derive the OLS estimator of α along with its mean and variance.

   (b) Derive the GLS estimator of α along with its mean and variance.

   (c) Obtain the relative efficiency of OLS with respect to GLS. Compute their relative efficiency for various values of $\sigma_2^2/\sigma_1^2=0.2,0.4,0.6,0.8,1,1.25,1.33,2.5,5$; and $n_1/n=0.2,0.3,0.4,\ldots,0.8$. Plot this relative efficiency.

   (d) Assume that u_i is $N(0,\sigma_1^2)$ for $i=1,2,\ldots,n_1$; and $N(0,\sigma_1^2)$ for $i=n_1+1,\ldots,n_1+n_2$; with u_i's being independent. What is the maximum likelihood estimator of α,σ_1^2 and σ_2^2?

   (e) Derive the LR test for testing $H_0;\sigma_1^2=\sigma_2^2$ in part (d).

6. Show that for an AR(1) model given in (5.26), $E(u_t u_s) = \rho^{|t-s|}\sigma_u^2$ for $t, s = 1, 2, \ldots, T$.

7. *Relative Efficiency of OLS Under the AR(1) Model.* This problem is based on Johnston (1984, pp. 310-312). For the simple regression without a constant $y_t = \beta x_t + u_t$ with $u_t = \rho u_{t-1} + \epsilon_t$ and $\epsilon_t \sim \text{IID}(0, \sigma_\epsilon^2)$

   (a) Show that

$$\begin{aligned}\text{var}(\widehat{\beta}_{OLS}) &= \frac{\sigma_u^2}{\sum_{t=1}^T x_t^2}\left(1 + 2\rho\frac{\sum_{t=1}^{T-1} x_t x_{t+1}}{\sum_{t=1}^T x_t^2} + 2\rho^2\frac{\sum_{t=1}^{T-2} x_t x_{t+2}}{\sum_{t=1}^T x_t^2}\right. \\ &\quad \left. + \ldots + 2\rho^{T-1}\frac{x_1 x_T}{\sum_{t=1}^T x_t^2}\right)\end{aligned}$$

   and that the Prais-Winsten estimator $\widehat{\beta}_{PW}$ has variance

$$\text{var}(\widehat{\beta}_{PW}) = \frac{\sigma_u^2}{\sum_{t=1}^T x_t^2}\left[\frac{1-\rho^2}{1+\rho^2 - 2\rho\sum_{t=1}^{T-1} x_t x_{t+1}/\sum_{t=1}^T x_t^2}\right]$$

   These expressions are easier to prove using matrix algebra, see Chapter 9.

   (b) Let x_t itself follow an AR(1) scheme with parameter λ, i.e., $x_t = \lambda x_{t-1} + v_t$, and let $T \to \infty$. Show that

$$\begin{aligned}\text{asy eff}(\widehat{\beta}_{OLS}) &= \lim_{T\to\infty}\frac{\text{var}(\widehat{\beta}_{PW})}{\text{var}(\widehat{\beta}_{OLS})} = \frac{1-\rho^2}{(1+\rho^2-2\rho\lambda)(1+2\rho\lambda+2\rho^2\lambda^2+\ldots)} \\ &= \frac{(1-\rho^2)(1-\rho\lambda)}{(1+\rho^2-2\rho\lambda)(1+\rho\lambda)}\end{aligned}$$

   (c) Tabulate this asy eff$(\widehat{\beta}_{OLS})$ for various values of ρ and λ where ρ varies between -0.9 to $+0.9$ in increments of 0.1, while λ varies between 0 and 0.9 in increments of 0.1. What do you conclude? How serious is the loss in efficiency in using OLS rather than the PW procedure?

   (d) Ignoring this autocorrelation one would compute $\sigma_u^2/\sum_{t=1}^T x_t^2$ as the var$(\widehat{\beta}_{OLS})$. The difference between this wrong formula and that derived in part (a) gives us the bias in estimating the variance of $\widehat{\beta}_{OLS}$. Show that as $T \to \infty$, this asymptotic proportionate bias is given by $-2\rho\lambda/(1+\rho\lambda)$. Tabulate this asymptotic bias for various values of ρ and λ as in part (c). What do you conclude? How serious is the asymptotic bias of using the wrong variances for $\widehat{\beta}_{OLS}$ when the disturbances are first-order autocorrelated?

   (e) Show that

$$\begin{aligned}E(s^2) &= \sigma_u^2\left\{T - \left(1 + 2\rho\frac{\sum_{t=1}^{T-1} x_t x_{t+1}}{\sum_{t=1}^T x_t^2} + 2\rho^2\frac{\sum_{t=1}^{T-2} x_t x_{t+2}}{\sum_{t=1}^T x_t^2}\right.\right. \\ &\quad \left.\left. + \ldots + 2\rho^{T-1}\frac{x_1 x_T}{\sum_{t=1}^T x_t^2}\right)\right\}/(T-1)\end{aligned}$$

   Conclude that if $\rho = 0$, then $E(s^2) = \sigma_u^2$. If x_t follows an AR(1) scheme with parameter λ, then for a large T, we get

$$E(s^2) = \sigma_u^2\left(T - \frac{1+\rho\lambda}{1-\rho\lambda}\right)/(T-1)$$

   Compute this $E(s^2)$ for $T = 101$ and various values of ρ and λ as in part (c). What do you conclude? How serious is the bias in using s^2 as an unbiased estimator for σ_u^2?

8. For the AR(1) model given in (5.26), show that if $\rho > 0$ and the x_t's are positively autocorrelated that $E(s^2/\sum x_t^2)$ understates the var$(\widehat{\beta}_{OLS})$ given in (5.34).

9. Show that for the AR(1) model, the Durbin-Watson statistic has plim$d \to 2(1-\rho)$.

10. *Regressions with Non-Zero Mean Disturbances.* Consider the simple regression with a constant

$$Y_i = \alpha + \beta X_i + u_i \quad i = 1, 2, \ldots, n$$

where α and β are scalars and u_i is independent of the X_i's. Show that:

   (a) If the u_i's are independent and identically gamma distributed with $f(u_i) = \frac{1}{\Gamma(\theta)} u_i^{\theta-1} e^{-u_i}$ where $u_i \geq 0$ and $\theta > 0$, then $\widehat{\alpha}_{OLS} - s^2$ is unbiased for α.

   (b) If the u_i's are independent and identically χ^2 distributed with ν degrees of freedom, then $\widehat{\alpha}_{OLS} - s^2/2$ is unbiased for α.

   (c) If the u_i's are independent and identically exponentially distributed with $f(u_i) = \frac{1}{\theta} e^{-u_i/\theta}$ where $u_i \geq 0$ and $\theta > 0$, then $\widehat{\alpha}_{OLS} - s$ is consistent for α.

11. *The Heteroskedastic Consequences of an Arbitrary Variance for the Initial Disturbance of an AR(1) Model.* This is based on Baltagi and Li (1990, 1992). Consider a simple AR(1) model

$$u_t = \rho u_{t-1} + \epsilon_t \quad t = 1, 2, \ldots, T \quad |\rho| < 1$$

with $\epsilon_t \sim \text{IID}(0, \sigma_\epsilon^2)$ independent of $u_0 \sim (0, \sigma_\epsilon^2/\tau)$, and τ is an arbitrary positive parameter.

   (a) Show that this arbitrary variance on the initial disturbance u_0 renders the disturbances, in general, heteroskedastic.

   (b) Show that var$(u_t) = \sigma_t^2$ is increasing if $\tau > (1-\rho^2)$ and decreasing if $\tau < (1-\rho^2)$. When is the process homoskedastic?

   (c) Show that cov$(u_t, u_{t-s}) = \rho^s \sigma_{t-s}^2$ for $t \geq s$. **Hint:** See the solution by Kim (1991).

   (d) Consider the simple regression model

$$y_t = \beta x_t + u_t \quad t = 1, 2 \ldots, T$$

   with u_t following the AR(1) process described above. Consider the common case where $\rho > 0$ and the x_t's are positively autocorrelated. For this case, it is a standard result that the var$(\widehat{\beta}_{OLS})$ is understated under the stationary case (i.e., $(1-\rho^2) = \tau$), see problem 8. This means that OLS rejects too often the hypothesis $H_0; \beta = 0$. Show that OLS will reject more often than the stationary case if $\tau < 1 - \rho^2$ and less often than the stationary case if $\tau > (1-\rho^2)$. **Hint:** See the solution by Koning (1992).

12. *ML Estimation of Linear Regression Model with AR(1) Errors and Two Observations.* This is based on Magee (1993). Consider the regression model $y_i = x_i\beta + u_i$, with only two observations $i = 1, 2$, and the nonstochastic $|x_1| \neq |x_2|$ are scalars. Assume that $u_i \sim N(0, \sigma^2)$ and $u_2 = \rho u_1 + \epsilon$ with $|\rho| < 1$. Also, $\epsilon \sim N[0, (1-\rho^2)\sigma^2]$ where ϵ and u_1 are independent.

   (a) Show that the OLS estimator of β is $(x_1y_1 + x_2y_2)/(x_1^2 + x_2^2)$.

   (b) Show that the ML estimator of β is $(x_1y_1 - x_2y_2)/(x_1^2 - x_2^2)$.

   (c) Show that the ML estimator of ρ is $2x_1x_2/(x_1^2 + x_2^2)$ and thus is nonstochastic.

   (d) How do the ML estimates of β and ρ behave as $x_1 \to x_2$ and $x_1 \to -x_2$? Assume $x_2 \neq 0$. **Hint:** See the solution by Baltagi and Li (1995).

13. For the empirical example in section 5.5 based on the Cigarette Consumption Data in Table 3.2.

(a) Replicate the OLS regression of $\log C$ on $\log P$, $\log Y$ and a constant. Plot the residuals versus $\log Y$ and verify Figure 3.2.

(b) Run Glejser's (1969) test by regressing $|e_i|$ the absolute value of the residuals from part (a), on $(\log Y_i)^{\delta}$ for $\delta = 1, -1, -0.5$ and 0.5. Verify the t-statistics reported in the text.

(c) Run Goldfeld and Quandt's (1965) test by ordering the observations according to $\log Y_i$ and omitting 12 central observations. Report the two regressions based on the first and last 17 observations and verify the F-test reported in the text.

(d) Verify the Spearman rank correlation test based on the rank $(\log Y_i)$ and rank $|e_i|$.

(e) Verify Harvey's (1976) multiplicative heteroskedasticity test based on regressing $\log e_i^2$ on $\log(\log Y_i)$.

(f) Run the Breusch and Pagan (1979) test based on the regression of $e_i^2/\widehat{\sigma}^2$ on $\log Y_i$, where $\widehat{\sigma}^2 = \sum_{i=1}^{46} e_i^2/46$.

(g) Run White's (1980) test for heteroskedasticity.

(h) Run the Jarque and Bera (1987) test for normality of the disturbances.

(i) Compute White's (1980) heteroskedasticity robust standard errors for the regression in part (a).

14. *A Simple Linear Trend Model with AR(1) Disturbances.* This is based on Krämer (1982).

(a) Consider the following simple linear trend model

$$Y_t = \alpha + \beta_t + u_t$$

where $u_t = \rho u_{t-1} + \epsilon_t$ with $|\rho| < 1$, $\epsilon_t \sim \text{IID}(0, \sigma_\epsilon^2)$ and $\text{var}(u_t) = \sigma_u^2 = \sigma_\epsilon^2/(1-\rho^2)$. Our interest is focused on the estimates of the trend coefficient, β, and the estimators to be considered are OLS, CO (assuming that the true value of ρ is known), the first-difference estimator (FD), and the Generalized Least Squares (GLS), which is Best Linear Unbiased (BLUE) in this case.

In the context of the simple linear trend model, the formulas for the variances of these estimators reduce to

$$\begin{aligned}
V(OLS) &= 12\sigma^2\{-6\rho^{T+1}[(T-1)\rho-(T+1)]^2-(T^3-T)\rho^4 \\
&\quad +2(T^2-1)(T-3)\rho^3+12(T^2+1)\rho^2-2(T^2-1)(T+3)\rho \\
&\quad +(T^3-T)\}/(1-\rho^2)(1-\rho)^4(T^3-T)^2 \\
V(CO) &= 12\sigma^2(1-\rho)^2(T^3-3T^2+2T), \\
V(FD) &= 2\sigma^2(1-\rho^{T-1})/(1-\rho^2)(T-1)^2, \\
V(GLS) &= 12\sigma^2/(T-1)[(T-3)(T-2)\rho^2-2(T-3)(T-1)\rho+T(T+1)].
\end{aligned}$$

(b) Compute these variances and their relative efficiency with respect to the GLS estimator for $T = 10, 20, 30, 40$ and ρ between -0.9 and 0.9 in 0.1 increments.

(c) For a given T, show that the limit of $\text{var}(OLS)/\text{var}(CO)$ is zero as $\rho \to 1$. Prove that $\text{var}(FD)$ and $\text{var}(GLS)$ both tend in the limit to $\sigma_\epsilon^2/(T-1) < \infty$ as $\rho \to 1$. Conclude that $\text{var}(GLS)/\text{var}(FD)$ tend to 1 as $\rho \to 1$. Also, show that $\lim_{\rho \to 1}[\text{var}(GLS)/\text{var}(OLS)] = 5(T^2+T)/6(T^2+1) < 1$ provided $T > 3$.

(d) For a given ρ, show that $\text{var}(FD) = O(T^{-2})$ whereas the variance of the remaining estimators is $O(T^{-3})$. Conclude that $\lim_{T \to \infty}[\text{var}(FD)/\text{var}(CO)] = \infty$ for any given ρ.

15. Consider the empirical example in section 5.6, based on the Consumption-Income data in Table 5.3. Obtain this data set from the CONSUMP.DAT file on the Springer web site.

(a) Replicate the OLS regression of C_t on Y_t and a constant, and compute the Durbin-Watson statistic. Test $H_0; \rho = 0$ versus $H_1; \rho > 0$ at the 5% significance level.

(b) Perform the Cochrane-Orcutt procedure and verify the regression results in the text.

(c) Perform the two-step Prais-Winsten procedure and verify the regression results in the text. Iterate on the Prais-Winsten procedure.

(d) Perform the maximum likelihood procedure and verify the results in the text.

(e) Perform Durbin's regression and verify the results in the text.

(f) Test for first-order serial correlation using the Breusch and Godfrey test.

(g) Compute the Newey-West heteroskedasticity and autocorrelation-consistent standard errors for the least squares estimates in part (a).

16. Benderly and Zwick (1985) considered the following equation

$$RS_t = \alpha + \beta Q_{t+1} + \gamma P_t + u_t$$

where RS_t = the real return on stocks in year t, Q_{t+1} = the annual rate of growth of real GNP in year $t+1$, and P_t = the rate of inflation in year t. The data is provided on the Springer web site and labeled BENDERLY.ASC. This data covers 31 annual observations for the U.S. over the period 1952-1982. This was obtained from Lott and Ray (1991). This equation is used to test the significance of the inflation rate in explaining real stock returns. Use the sample period 1954-1976 to answer the following questions:

(a) Run OLS to estimate the above equation. Remember to use Q_{t+1}. Is P_t significant in this equation? Plot the residuals against time. Compute the Newey-West heteroskedasticity and autocorrelation-consistent standard errors for these least squares estimates.

(b) Test for serial correlation using the D.W. test.

(c) Would your decision in (b) change if you used the Breusch-Godfrey test for first-order serial correlation?

(d) Run the Cochrane-Orcutt procedure to correct for first-order serial correlation. Report your estimate of ρ.

(e) Run a Prais-Winsten procedure accounting for the first observation and report your estimate of ρ. Plot the residuals against time.

17. Using our cross-section Energy/GDP data set in Chapter 3, problem 3.16 consider the following two models:

$$\begin{aligned} &Model\ 1: \quad \log En = \alpha + \beta \log RGDP + u \\ &Model\ 2: \quad En = \alpha + \beta RGDP + v \end{aligned}$$

Make sure you have corrected the W. Germany observation on EN as described in problem 3.16 part (d).

(a) Run OLS on both Models 1 and 2. Test for heteroskedasticity using the Goldfeldt/Quandt Test. Omit $c = 6$ central observations. Why is heteroskedasticity a problem in Model 2, but not Model 1?

(b) For Model 2, test for heteroskedasticity using the Glejser Test.

(c) Now use the Breusch-Pagan Test to test for heteroskedasticity on Model 2.

(d) Apply White's Test to Model 2.

(e) Do all these tests give the same decision?

(f) Propose and estimate a simple transformation of Model 2, assuming heteroskedasticity of the form $\sigma_i^2 = \sigma^2 RGDP^2$.

(g) Propose and estimate a simple transformation of Model 2, assuming heteroskedasticity of the form $\sigma_i^2 = \sigma^2(a + bRGDP)^2$.

(h) Now suppose that heteroskedasticity is of the form $\sigma_i^2 = \sigma^2 RGDP^\gamma$ where γ is an unknown parameter. Propose and estimate a simple transformation for Model 2. **Hint:** You can write σ_i^2 as $\exp\{\alpha + \gamma \log RGDP\}$ where $\alpha = \log\sigma^2$.

(i) Compare the standard errors of the estimates for Model 2 from OLS, also obtain White's heteroskedasticity-consistent standard errors. Compare them with the simple Weighted Least Squares estimates of the standard errors in parts (f), (g) and (h). What do you conclude?

18. You are given quarterly data from the first quarter of 1965 (1965.1) to the fourth quarter of 1983 (1983.4) on employment in Orange County California (EMP) and real gross national product (RGNP). The data set is in a file called ORANGE.DAT on the Springer web site.

(a) Generate the lagged variable of real GNP, call it $RGNP_{t-1}$ and estimate the following model by OLS: $EMP_t = \alpha + \beta RGNP_{t-1} + u_t$.

(b) What does inspection of the residuals and the Durbin-Watson statistic suggest?

(c) Assuming $u_t = \rho u_{t-1} + \epsilon_t$ where $|\rho| < 1$ and $\epsilon_t \sim \text{IIN}(0, \sigma_\epsilon^2)$, use the Cochrane-Orcutt procedure to estimate ρ, α and β. Compare the latter estimates and their standard errors with those of OLS.

(d) The Cochrane-Orcutt procedure omits the first observation. Perform the Prais-Winsten adjustment. Compare the resulting estimates and standard error with those in part (c).

(e) Apply the Breusch-Godfrey test for first and second order autoregression. What do you conclude?

(f) Compute the Newey-West heteroskedasticity and autocorrelation-consistent covariance standard errors for the least squares estimates in part (a).

19. Consider the earning data underlying the regression in Table 4.1 and available on the Springer web site as EARN.ASC.

(a) Apply White's test for heteroskedasticity to the regression residuals.

(b) Compute White's heteroskedasticity-consistent standard errors.

(c) Test the least squares residuals for normality using the Jarque-Bera test.

20. Harrison and Rubinfield (1978) collected data on 506 census tracts in the Boston area in 1970 to study hedonic housing prices and the willingness to pay for clean air. This data is available on the Springer web site as HEDONIC.XLS. The dependent variable is the *Median Value* (MV) of owner-occupied homes. The regressors include two structural variables, RM the average number of rooms, and AGE representing the proportion of owner units built prior to 1940. In addition there are eight neighborhood variables: B, the proportion of blacks in the population; LSTAT, the proportion of population that is lower status; CRIM, the crime rate; ZN, the proportion of 25000 square feet residential lots; INDUS, the proportion of nonretail business acres; TAX, the full value property tax rate (\$/\$10000); PTRATIO, the pupil-teacher ratio; and CHAS represents the dummy variable for Charles River: = 1 if a tract bounds the Charles. There are also two accessibility variables, DIS the weighted distances to five employment centers in the Boston region, and RAD the index of accessibility to radial highways. One more regressor is an air pollution variable NOX, the annual average nitrogen oxide concentration in parts per hundred million.

(a) Run OLS of MV on the 13 independent variables and a constant. Plot the residuals.

(b) Apply White's tests for heteroskedasticity.

(c) Obtain the White heteroskedasticity-consistent standard errors.

(d) Test the least squares residuals for normality using the Jarque-Bera test.

References

For additional readings consult the econometrics books cited in the Preface. Recent chapters on heteroskedasticity and autocorrelation include Griffiths (2001) and King (2001):

Ali, M.M. and C. Giaccotto (1984), "A study of Several New and Existing Tests for Heteroskedasticity in the General Linear Model," *Journal of Econometrics*, 26: 355-373.

Amemiya, T. (1973), "Regression Analysis When the Variance of the Dependent Variable is Proportional to the Square of its Expectation," *Journal of the American Statistical Association*, 68: 928-934.

Amemiya, T. (1977), "A Note on a Heteroskedastic Model," *Journal of Econometrics*, 6: 365-370.

Andrews, D.W.K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59: 817-858.

Baltagi, B. and Q. Li (1990), "The Heteroskedastic Consequences of an Arbitrary Variance for the Initial Disturbance of an AR(1) Model," *Econometric Theory*, Problem 90.3.1, 6: 405.

Baltagi, B. and Q. Li (1992), "The Bias of the Standard Errors of OLS for an AR(1) process with an Arbitrary Variance on the Initial Observations," *Econometric Theory*, Problem 92.1.4, 8: 146.

Baltagi, B. and Q. Li (1995), "ML Estimation of Linear Regression Model with AR(1) Errors and Two Observations," *Econometric Theory*, Solution 93.3.2, 11: 641-642.

Bartlett's test, M.S. (1937), "Properties of Sufficiency and Statistical Tests," *Proceedings of the Royal Statistical Society*, A, 160: 268-282.

Beach, C.M. and J.G. MacKinnon (1978), "A Maximum Likelihood Procedure for Regression with Autocorrelated Errors," *Econometrica*, 46: 51-58.

Benderly, J. and B. Zwick (1985), "Inflation, Real Balances, Output and Real Stock Returns," *American Economic Review*, 75: 1115-1123.

Breusch, T.S. (1978), "Testing for Autocorrelation in Dynamic Linear Models," *Australian Economic Papers*, 17: 334-355.

Breusch, T.S. and A.R. Pagan (1979), "A Simple Test for Heteroskedasticity and Random Coefficient Variation," *Econometrica*, 47: 1287-1294.

Buse, A. (1984), "Tests for Additive Heteroskedasticity: Goldfeld and Quandt Revisited," *Empirical Economics*, 9: 199-216.

Carroll, R.H. (1982), "Adapting for Heteroskedasticity in Linear Models," *Annals of Statistics*, 10: 1224-1233.

Cochrane, D. and G. Orcutt (1949), "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms," *Journal of the American Statistical Association*, 44: 32-61.

Cragg, J.G. (1992), "Quasi-Aitken Estimation for Heteroskedasticity of Unknown Form," *Journal of Econometrics*, 54: 197-202.

Durbin, J. (1960), "Estimation of Parameters in Time-Series Regression Model," *Journal of the Royal Statistical Society*, Series B, 22: 139-153.

Durbin, J. and G. Watson (1950), "Testing for Serial Correlation in Least Squares Regression-I," *Biometrika*, 37: 409-428.

Durbin, J. and G. Watson (1951), "Testing for Serial Correlation in Least Squares Regression-II," *Biometrika*, 38: 159-178.

Evans, M.A., and M.L. King (1980) "A Further Class of Tests for Heteroskedasticity," *Journal of Econometrics*, 37: 265-276.

Farebrother, R.W. (1980), "The Durbin-Watson Test for Serial Correlation When There is No Intercept in the Regression," *Econometrica*, 48: 1553-1563.

Glejser, H. (1969), "A New Test for Heteroskedasticity," *Journal of the American Statistical Association*, 64: 316-323.

Godfrey, L.G. (1978), "Testing Against General Autoregressive and Moving Average Error Models When the Regressors Include Lagged Dependent Variables," *Econometrica*, 46: 1293-1302.

Goldfeld, S.M. and R.E. Quandt (1965), "Some Tests for Homoscedasticity," *Journal of the American Statistical Association*, 60: 539-547.

Goldfeld, S.M. and R.E. Quandt (1972), *Nonlinear Methods in Econometrics* (North-Holland: Amsterdam).

Griffiths, W.E. (2001), "Heteroskedasticity," Chapter 4 in B.H. Baltagi, (ed.), *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Harrison, M. and B.P. McCabe (1979), "A Test for Heteroskedasticity Based On Ordinary Least Squares Residuals," *Journal of the American Statistical Association*, 74: 494-499.

Harrison, D. and D.L. Rubinfeld (1978), "Hedonic Housing Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management*, 5: 81-102.

Harvey, A.C. (1976), "Estimating Regression Models With Multiplicative Heteroskedasticity," *Econometrica*, 44: 461-466.

Hilderth, C. and J. Lu (1960), "Demand Relations with Autocorrelated Disturbances," Technical Bulletin 276 (Michigan State University, Agriculture Experiment Station).

Jarque, C.M. and A.K. Bera (1987), "A Test for Normality of Observations and Regression Residuals," *International Statistical Review*, 55: 163-177.

Kim, J.H. (1991), "The Heteroskedastic Consequences of an Arbitrary Variance for the Initial Disturbance of an AR(1) Model," *Econometric Theory*, Solution 90.3.1, 7: 544-545.

King, M. (2001), "Serial Correlation," Chapter 2 in B.H. Baltagi, (ed.), *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Koenker, R. (1981), "A Note on Studentizing a Test for Heteroskedasticity," *Journal of Econometrics*, 17: 107-112.

Koenker, R. and G.W. Bassett, Jr. (1982), "Robust Tests for Heteroskedasticity Based on Regression Quantiles," *Econometrica*, 50:43-61.

Koning, R.H. (1992), "The Bias of the Standard Errors of OLS for an AR(1) process with an Arbitrary Variance on the Initial Observations," *Econometric Theory*, Solution 92.1.4, 9: 149-150.

Krämer, W. (1982), "Note on Estimating Linear Trend When Residuals are Autocorrelated," *Econometrica*, 50: 1065-1067.

Lott, W.F. and S.C. Ray (1992), *Applied Econometrics: Problems With Data Sets* (The Dryden Press: New York).

Maddala, G.S. (1977), *Econometrics* (McGraw-Hill: New York).

Maeshiro, A. (1976), "Autoregressive Transformation, Trended Independent Variables and Autocorrelated Disturbance Terms," *The Review of Economics and Statistics*, 58: 497-500.

Maeshiro, A. (1979), "On the Retention of the First Observations in Serial Correlation Adjustment of Regression Models," *International Economic Review*, 20: 259-265.

Magee L. (1993), "ML Estimation of Linear Regression Model with AR(1) Errors and Two Observations," *Econometric Theory*, Problem 93.3.2, 9: 521-522.

Mizon, G.E. (1995), "A Simple Message for Autocorrelation Correctors: Don't," *Journal of Econometrics* 69: 267-288.

Newey, W.K. and K.D. West (1987), "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55: 703-708.

Oberhofer, W. and J. Kmenta (1974), "A General Procedure for Obtaining Maximum Likelihood Estimates in Generalized Regression Models," *Econometrica*, 42: 579-590.

Park, R.E. and B.M. Mitchell (1980), "Estimating the Autocorrelated Error Model With Trended Data," *Journal of Econometrics*, 13: 185-201.

Prais, S. and C. Winsten (1954), "Trend Estimation and Serial Correlation," Discussion Paper 383 (Cowles Commission: Chicago).

Rao, P. and Z. Griliches (1969), "Some Small Sample Properties of Several Two-Stage Regression Methods in the Context of Autocorrelated Errors," *Journal of the American Statistical Association*, 64: 253-272.

Rao, P. and R. L. Miller (1971), *Applied Econometrics* (Wadsworth: Belmont).

Robinson, P.M. (1987), "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form," *Econometrica*, 55: 875-891.

Rutemiller, H.C. and D.A. Bowers (1968), "Estimation in a Heteroskedastic Regression Model," *Journal of the American Statistical Association*, 63: 552-557.

Savin, N.E. and K.J. White (1977), "The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors," *Econometrica*, 45: 1989-1996.

Szroeter, J. (1978), "A Class of Parametric Tests for Heteroskedasticity in Linear Econometric Models," *Econometrica*, 46: 1311-1327.

Theil, H. (1978), *Introduction to Econometrics* (Prentice-Hall: Englewood Cliffs, NJ).

Waldman, D.M. (1983), "A Note on Algebraic Equivalence of White's Test and a Variation of the Godfrey/Breusch-Pagan Test for Heteroskedasticity," *Economics Letters*, 13: 197-200.

White, H. (1980), "A Heteroskedasticity Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48: 817-838.

Wooldridge, J.M. (1991), "On the Application of Robust, Regression-Based Diagnostics to Models of Conditional Means and Conditional Variances," *Journal of Econometrics*, 47: 5-46.

CHAPTER 6
Distributed Lags and Dynamic Models

6.1 Introduction

Many economic models have lagged values of the regressors in the regression equation. For example, it takes time to build roads and highways. Therefore, the effect of this public investment on growth in GNP will show up with a lag, and this effect will probably linger on for several years. It takes time before investment in research and development pays off in new inventions which in turn take time to develop into commercial products. In studying consumption behavior, a change in income may affect consumption over several periods. This is true in the permanent income theory of consumption, where it may take the consumer several periods to determine whether the change in real disposable income was temporary or permanent. For example, is the extra consulting money earned this year going to continue next year? Also, lagged values of real disposable income appear in the regression equation because the consumer takes into account his life time earnings in trying to smooth out his consumption behavior. In turn, one's life time income may be guessed by looking at past as well as current earnings. In other words, the regression relationship would look like

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + .. + \beta_s X_{t-s} + u_t \quad t = 1, 2, \ldots, T \tag{6.1}$$

where Y_t denotes the t-th observation on the dependent variable Y and X_{t-s} denotes the $(t\text{-}s)$th observation on the independent variable X. α is the intercept and $\beta_0, \beta_1, \ldots, \beta_s$ are the current and lagged coefficients of X_t. Equation (6.1) is known as a *distributed lag* since it distributes the effect of an increase in income on consumption over s periods. Note that the *short-run* effect of a unit change in X on Y is given by β_o, while the *long-run* effect of a unit change in X on Y is $(\beta_0 + \beta_1 + .. + \beta_s)$.

Suppose that we observe X_t from 1955 to 1995. X_{t-1} is the same variable but for the previous period, i.e., 1954-1994. Since 1954 is not observed, we start from 1955 for X_{t-1}, and end at 1994. This means that when we lag once, the current X_t series will have to start at 1956 and end at 1995. For practical purposes, this means that when we lag once we loose one observation from the sample. So if we lag s periods, we loose s observations. Furthermore, we are estimating one extra β with every lag. Therefore, there is double jeopardy with respect to loss of degrees of freedom. The number of observations fall (because we are lagging the same series), and the number of parameters to be estimated increase with every lagging variable introduced. Besides the loss of degrees of freedom, the regressors in (6.1) are likely to be highly correlated with each other. In fact most economic time series are usually trended and very highly correlated with their lagged values. This introduces the problem of multicollinearity among the regressors and as we saw in Chapter 4, the higher the multicollinearity among these regressors, the lower is the reliability of the regression estimates.

In this model, OLS is still BLUE because the classical assumptions are still satisfied. All we have done in (6.1) is introduce the additional regressors $(X_{t-1}, \ldots, X_{t-s})$. These regressors are uncorrelated with the disturbances since they are lagged values of X_t, which are by assumption not correlated with u_t for every t.

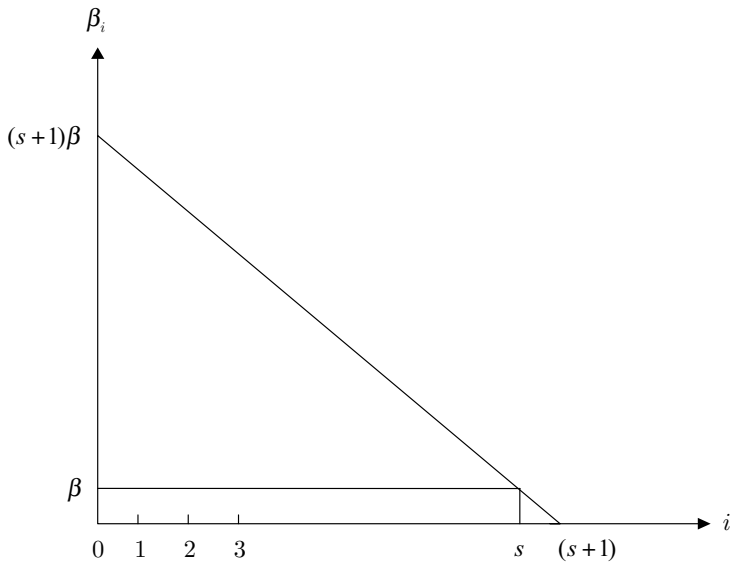


Figure 6.1 Linear Distributed Lag

In order to reduce the degrees of freedom problem, one could impose more structure on the β's. One of the simplest forms imposed on these coefficients is the *linear arithmetic lag*, (see Figure 6.1), which can be written as

$$\beta_i = [(s+1) - i]\beta \quad \text{for } i = 0, 1, \ldots, s \tag{6.2}$$

The lagged coefficients of X follow a linear distributed lag declining arithmetically from $(s+1)\beta$ for X_t to β for X_{t-s}. Substituting (6.2) in (6.1) one gets

$$Y_t = \alpha + \textstyle\sum_{i=0}^{s} \beta_i X_{t-i} + u_t = \alpha + \beta \sum_{i=0}^{s} [(s+1) - i] X_{t-i} + u_t \tag{6.3}$$

where the latter equation can be estimated by the regression of Y_t on a constant and Z_t, where

$$Z_t = \textstyle\sum_{i=0}^{s} [(s+1) - i] X_{t-i}$$

This Z_t can be calculated given s and X_t. Hence, we have reduced the estimation of $\beta_0, \beta_1, \ldots, \beta_s$ into the estimation of just one β. Once $\widehat{\beta}$ is obtained, $\widehat{\beta}_i$ can be deduced from (6.2), for $i = 0, 1, \ldots, s$. Despite its simplicity, this lag is too restrictive to impose on the regression and is not usually used in practice.

Alternatively, one can think of $\beta_i = f(i)$ for $i = 0, 1, \ldots, s$. If $f(i)$ is a continuous function, over a closed interval, then it can be approximated by an r-th degree polynomial,

$$f(i) = a_0 + a_1 i + \ldots + a_r i^r$$

For example, if $r = 2$, then

$$\beta_i = a_0 + a_1 i + a_2 i^2 \quad \text{for } i = 0, 1, 2, \ldots, s$$

so that

$$\begin{aligned}
\beta_0 &= a_0 \\
\beta_1 &= a_0 + a_1 + a_2 \\
\beta_2 &= a_0 + 2a_1 + 4a_2 \\
&\vdots \quad \vdots \quad \vdots \\
\beta_s &= a_0 + sa_1 + s^2 a_2
\end{aligned}$$

Once a_0, a_1, and a_2 are estimated, $\beta_0, \beta_1, \ldots, \beta_s$ can be deduced. In fact, substituting $\beta_i = a_0 + a_1 i + a_2 i^2$ in (6.1) we get

$$\begin{aligned}
Y_t &= \alpha + \textstyle\sum_{i=0}^{s}(a_0 + a_1 i + a_2 i^2)X_{t-i} + u_t \\
&= \alpha + a_0 \textstyle\sum_{i=0}^{s} X_{t-i} + a_1 \sum_{i=0}^{s} iX_{t-i} + a_2 \sum_{i=0}^{s} i^2 X_{t-i} + u_t
\end{aligned} \tag{6.4}$$

This last equation, shows that α, a_0, a_1 and a_2 can be estimated from the regression of Y_t on a constant, $Z_0 = \sum_{i=0}^{s} X_{t-i}$, $Z_1 = \sum_{i=0}^{s} iX_{t-i}$ and $Z_2 = \sum_{i=0}^{s} i^2 X_{t-i}$. This procedure was proposed by Almon (1965) and is known as the *Almon lag*. One of the problems with this procedure is the choice of s and r, the number of lags on X_t, and the degree of the polynomial, respectively. In practice, neither is known. Davidson and MacKinnon (1993) suggest starting with a maximum reasonable lag s^* that is consistent with the theory and then based on the unrestricted regression, given in (6.1), checking whether the fit of the model deteriorates as s^* is reduced. Some criteria suggested for this choice include: (i) maximizing $\bar{R}^2$; (ii) minimizing *Akaike's* (1973) *Information Criterion* (AIC) with respect to s. This is given by $AIC(s) = (RSS/T)e^{2s/T}$; or (iii) minimizing Schwarz (1978) *Bayesian Information Criterion* (BIC) with respect to s. This is given by $BIC(s) = (RSS/T)T^{s/T}$ where RSS denotes the residual sum of squares. Note that the AIC and BIC criteria, like $\bar{R}^2$, reward good fit but penalize loss of degrees of freedom associated with a high value of s. These criteria are printed by most regression software including SHAZAM, EViews and SAS. Once the lag length s is chosen it is straight forward to determine r, the degree of the polynomial. Start with a high value of r and construct the Z variables as described in (6.4). If $r = 4$ is the highest degree polynomial chosen and a_4, the coefficient of $Z_4 = \sum_{i=0}^{s} i^4 X_{t-4}$ is insignificant, drop Z_4 and run the regression for $r = 3$. Stop, if the coefficient of Z_3 is significant, otherwise drop Z_3 and run the regression for $r = 2$.

Applied researchers usually impose end point constraints on this Almon lag. A near end point constraint means that $\beta_{-1} = 0$ in equation (6.1). This means that for equation (6.4), this constraint yields the following restriction on the second degree polynomial in a's: $\beta_{-1} = f(-1) = a_0 - a_1 + a_2 = 0$. This restriction allows us to solve for a_0 given a_1 and a_2. In fact, substituting $a_0 = a_1 - a_2$ into (6.4), the regression becomes

$$Y_t = \alpha + a_1(Z_1 + Z_0) + a_2(Z_2 - Z_0) + u_t \tag{6.5}$$

and once a_1 and a_2 are estimated, a_0 is deduced, and hence the β_i's. This restriction essentially states that X_{t+1} has no effect on Y_t. This may not be a plausible assumption, especially in our consumption example, where income next year enters the calculation of permanent income or life time earnings. A more plausible assumption is the far end point constraint, where $\beta_{s+1} = 0$. This means that $X_{t-(s+1)}$ does not affect Y_t. The further you go back in time, the less is the effect on the current period. All we have to be sure of is that we have gone far back enough

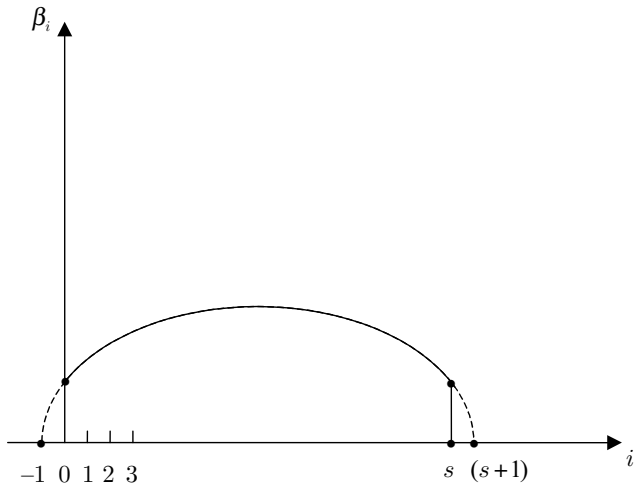


Figure 6.2 A Polynomial Lag with End Point Constraints

to reach an insignificant effect. This far end point constraint is imposed by removing $X_{t-(s+1)}$ from the equation as we have done above. But, some researchers impose this restriction on $\beta_i = f(i)$, i.e., by restricting $\beta_{s+1} = f(s+1) = 0$. This yields for $r = 2$ the following constraint: $a_0+(s+1)a_1+(s+1)^2a_2 = 0$. Solving for a_0 and substituting in (6.4), the constrained regression becomes

$$Y_t = \alpha + a_1[Z_1 - (s+1)Z_0] + a_2[Z_2 - (s+1)^2 Z_0] + u_t \tag{6.6}$$

One can also impose both end point constraints and reduce the regression into the estimation of one a rather than three a's. Note that $\beta_{-1} = \beta_{s+1} = 0$ can be imposed by not including X_{t+1} and $X_{t-(s+1)}$ in the regression relationship. However, these end point restrictions impose the *additional* restrictions that the polynomial on which the a's lie should pass through zero at $i = -1$ and $i = (s+1)$, see Figure 6.2.

These additional restrictions on the polynomial may not necessarily be true. In other words, the polynomial could intersect the X-axis at points other than -1 or $(s+1)$. Imposing a restriction, whether true or not, reduces the variance of the estimates, and introduces bias if the restriction is untrue. This is intuitive, because this restriction gives *additional* information which should increase the reliability of the estimates. The reduction in variance and the introduction of bias naturally lead to Mean Square Error criteria that help determine whether these restrictions should be imposed, see Wallace (1972). These criteria are beyond the scope of this chapter. In general, one should be careful in the use of restrictions that may not be plausible or even valid. In fact, one should always test these restrictions before using them. See Schmidt and Waud (1975).

Empirical Example: Using the Consumption-Income data from the Economic Report of the President over the period 1950-1993, given in Table 5.1, we estimate a consumption-income regression imposing a five year lag on income. In this case, all variables are in logs and $s = 5$ in equation (6.1). Table 6.1 gives the SAS output imposing the *linear arithmetic* lag given in equation (6.2).

Note that the SAS output reports $\widehat{\beta} = 0.047$ as the coefficient of Y_{t-5} which is denoted by YLAG5. This is statistically significant with a t-value of 83.1. Note that the coefficient of Y_{t-4}

Table 6.1 Regression with Arithmetic Lag Restriction

Dependent Variable: C
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	2.30559	2.30559	6902.791	0.0001
Error	37	0.01236	0.00033		
C Total	38	2.31794			

Root MSE	0.01828	R-square	0.9947
Dep Mean	9.14814	Adj R-sq	0.9945
C.V.	0.19978		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	0.095213	0.10900175	0.873	0.3880
Y	1	0.280795	0.00337969	83.083	0.0001
YLAG1	1	0.233996	0.00281641	83.083	0.0001
YLAG2	1	0.187197	0.00225313	83.083	0.0001
YLAG3	1	0.140398	0.00168985	83.083	0.0001
YLAG4	1	0.093598	0.00112656	83.083	0.0001
YLAG5	1	0.046799	0.00056328	83.083	0.0001
RESTRICT	–1	0.007218	0.00184780	3.906	0.0004
RESTRICT	–1	0.000781	0.00123799	0.631	0.5319
RESTRICT	–1	–0.003911	0.00127903	–3.058	0.0041
RESTRICT	–1	–0.005374	0.00188105	–2.857	0.0070
RESTRICT	–1	–0.005208	0.00261513	–1.991	0.0539

which is denoted by YLAG4 is $2\widehat{\beta}$, and so on. The coefficient of Y_t is given by $6\widehat{\beta} = 0.281$. At the bottom of the regression output, SAS tests each one of these five coefficient restrictions individually. We can see that three of these restrictions are rejected at the 5% level. One can test the arithmetic lag restrictions jointly using an F-test. The Unrestricted Residual Sum of Squares (URSS) is obtained by regressing C_t on $Y_t, Y_{t-1}, \ldots, Y_{t-5}$ and a constant. This yields $URSS = 0.00667$. The RRSS is given in Table 6.1 as 0.01236 and it involves imposing 5 restrictions given in (6.2). Therefore,

$$F = \frac{(0.01236 - 0.00667)/5}{0.00667/32} = 5.4597$$

and this is distributed as $F_{5,32}$ under the null hypothesis. The observed F-statistic has a p-value of 0.001 and we reject the linear arithmetic lag restrictions.

Next we impose an Almon lag based on a second degree polynomial as described in equation (6.4). Table 6.2 reports the SAS output for $s = 5$ imposing the near end point constraint. In this case, the estimated regression coefficients rise and then fall: $\widehat{\beta}_0 = 0.193$, $\widehat{\beta}_1 = 0.299, \ldots, \widehat{\beta}_5 = -0.159$. Only $\widehat{\beta}_5$ is statistically insignificant. In addition, SAS reports a t-test for the near end point restriction which is rejected with a p-value of 0.0001. The Almon lag restrictions can be

Table 6.2 Almon Polynomial, $r = 2, s = 5$ and Near End-Point Constraint

PDLREG Procedure
Dependent Variable = C
Ordinary Least Squares Estimates

SSE	0.014807	DFE	36
MSE	0.000411	Root MSE	0.020281
SBC	–185.504	AIC	–190.495
Reg Rsq	0.9936	Total Rsq	0.9936
Durbin-Watson	0.6958		

Variable	DF	B Value	Std Error	t Ratio	Approx Prob
Intercept	1	0.093289	0.1241	0.752	0.4571
Y**0	1	0.400930	0.00543	73.838	0.0001
Y**1	1	–0.294560	0.0892	–3.302	0.0022
Y**2	1	–0.268490	0.0492	–5.459	0.0001

Restriction	DF	L Value	Std Error	t Ratio	Approx Prob
Y(–1)	–1	0.005691	0.00135	4.203	0.0001

Variable	Parameter Value	Std Error	t Ratio	Approx Prob	Estimate of Lag Distribution −0.159 0 0.3161
Y(0)	0.19324	0.027	7.16	0.0001	| | **************** |
Y(1)	0.29859	0.038	7.88	0.0001	| | ************************ |
Y(2)	0.31606	0.033	9.67	0.0001	| | ****************************** |
Y(3)	0.24565	0.012	21.20	0.0001	| | ********************* |
Y(4)	0.08735	0.026	3.32	0.0020	| | ******* |
Y(5)	–0.15883	0.080	–1.99	0.0539	|*************| | |

jointly tested using Chow's F-statistic. The URSS is obtained from the unrestricted regression of C_t on $Y_t, Y_{t-1}, \ldots, Y_{t-5}$ and a constant. This was reported above as $URSS = 0.00667$.

The RRSS, given in Table 6.2, is 0.014807 and involves four restrictions. Therefore,

$$F = \frac{(0.014807 - 0.00667)/4}{0.00667/32} = 9.76$$

and this is distributed as $F_{4,32}$ under the null hypothesis. The observed F-statistic has a p-value of 0.00003 and we reject the second degree polynomial Almon lag specification with a near end point constraint.

Table 6.3 reports the SAS output for $s = 5$, imposing the far end point constraint. Note that this restriction is rejected with a p-value of 0.008. In this case, the $\widehat{\beta}$'s are decreasing, $\widehat{\beta}_0 = 0.502$, $\widehat{\beta}_1 = 0.309, \ldots, \widehat{\beta}_5 = -0.026$ with $\widehat{\beta}_4$ and $\widehat{\beta}_5$ being statistically insignificant. Most packages have polynomial distributed lags as part of their standard commands. For example, using EViews, replacing the regressor Y by PDL$(Y, 5, 2, 1)$ indicates the request to fit a five year Almon lag on Y that is of the second-order degree, with a near end point constraint.

Table 6.3 Almon Polynomial, $r = 2, s = 5$ and Far End-Point Constraint

PDLREG Procedure

Dependent Variable = C
Ordinary Least Squares Estimates

SSE	0.009244	DFE	36
MSE	0.000257	Root MSE	0.016024
SBC	–203.879	AIC	–208.87
Reg Rsq	0.9960	Total Rsq	0.9960
Durbin-Watson	0.6372		

Variable	DF	B Value	Std Error	t Ratio	Approx Prob
Intercept	1	–0.015868	0.1008	–0.157	0.8757
Y**0	1	0.405244	0.00439	92.331	0.0001
Y**1	1	–0.441447	0.0706	–6.255	0.0001
Y**2	1	–0.133484	0.0383	–3.483	0.0013

Restriction	DF	L Value	Std Error	t Ratio	Approx Prob
Y(–1)	–1	0.002758	0.00107	2.575	0.0080

Variable	Parameter Value	Std Error	t Ratio	Approx Prob	Estimate of Lag Distribution –0.026	0.5021
Y(0)	0.50208	0.064	7.89	0.0001		***************************************
Y(1)	0.30916	0.022	14.32	0.0001		************************
Y(2)	0.15995	0.008	19.82	0.0001		************
Y(3)	0.05442	0.025	2.20	0.0343		****
Y(4)	–0.00741	0.029	–0.26	0.7998		
Y(5)	–0.02555	0.021	–1.23	0.2268	*	

6.2 Infinite Distributed Lag

So far we have been dealing with a finite number of lags imposed on X_t. Some lags may be infinite. For example, the investment in building highways and roads several decades ago may still have an effect on today's growth in GNP. In this case, we write equation (6.1) as

$$Y_t = \alpha + \sum_{i=0}^{\infty} \beta_i X_{t-i} + u_t \quad t = 1, 2, \ldots, T. \tag{6.7}$$

There are an infinite number of β_i's to estimate with only T observations. This can only be feasible if more structure is imposed on the β_i's. First, we normalize these β_i's by their sum, i.e., let $w_i = \beta_i/\beta$ where $\beta = \sum_{i=0}^{\infty} \beta_i$. If all the β_i's have the *same sign*, then the β_i's take the sign of β and $0 \leq w_i \leq 1$ for all i, with $\sum_{i=0}^{\infty} w_i = 1$. This means that the w_i's can be interpreted as probabilities. In fact, Koyck (1954) imposed the geometric lag on the w_i's, i.e., $w_i = (1-\lambda)\lambda^i$ for $i = 0, 1, \ldots, \infty$[1]. Substituting

$$\beta_i = \beta w_i = \beta(1-\lambda)\lambda^i$$

in (6.7) we get

$$Y_t = \alpha + \beta(1-\lambda) \sum_{i=0}^{\infty} \lambda^i X_{t-i} + u_t \tag{6.8}$$

Equation (6.8) is known as the *infinite distributed* lag form of the *Koyck lag*. The short-run effect of a unit change in X_t on Y_t is given by $\beta(1-\lambda)$; whereas the long-run effect of a unit change in X_t on Y_t is $\sum_{i=0}^{\infty}\beta_i = \beta\sum_{i=0}^{\infty} w_i = \beta$. Implicit in the Koyck lag structure is that the effect of a unit change in X_t on Y_t declines the further back we go in time. For example, if $\lambda = 1/2$, then $\beta_0 = \beta/2$, $\beta_1 = \beta/4$, $\beta_2 = \beta/8$, etc. Defining $LX_t = X_{t-1}$, as the lag operator, we have $L^iX_t = X_{t-i}$, and (6.8) reduces to

$$Y_t = \alpha + \beta(1-\lambda)\sum_{i=0}^{\infty}(\lambda L)^i X_t + u_t = \alpha + \beta(1-\lambda)X_t/(1-\lambda L) + u_t \tag{6.9}$$

where we have used the fact that $\sum_{i=0}^{\infty} c^i = 1/(1-c)$. Multiplying the last equation by $(1-\lambda L)$ one gets

$$Y_t - \lambda Y_{t-1} = \alpha(1-\lambda) + \beta(1-\lambda)X_t + u_t - \lambda u_{t-1}$$

or

$$Y_t = \lambda Y_{t-1} + \alpha(1-\lambda) + \beta(1-\lambda)X_t + u_t - \lambda u_{t-1} \tag{6.10}$$

This is the *autoregressive form* of the infinite distributed lag. It is autoregressive because it includes the lagged value of Y_t as an explanatory variable. Note that we have reduced the problem of estimating an infinite number of β_i's into estimating λ and β from (6.10). However, OLS would lead to biased and inconsistent estimates, because (6.10) contains a lagged dependent variable as well as serially correlated errors. In fact the error in (6.10) is a Moving Average process of order one, i.e., MA(1), see Chapter 14. We digress at this stage to give two econometric models which would lead to equations resembling (6.10).

6.2.1 Adaptive Expectations Model (AEM)

Suppose that output Y_t is a function of expected sales X_t^* and that the latter is unobservable, i.e.,

$$Y_t = \alpha + \beta X_t^* + u_t$$

where expected sales are updated according to the following method

$$X_t^* - X_{t-1}^* = \delta(X_t - X_{t-1}^*) \tag{6.11}$$

that is, expected sales at time t is a weighted combination of expected sales at time $t-1$ and actual sales at time t. In fact,

$$X_t^* = \delta X_t + (1-\delta)X_{t-1}^* \tag{6.12}$$

Equation (6.11) is also an error learning model, where one learns from past experience and adjust expectations after observing current sales. Using the lag operator L, (6.12) can be rewritten as $X_t^* = \delta X_t/[1-(1-\delta)L]$. Substituting this last expression in the above relationship, we get

$$Y_t = \alpha + \beta\delta X_t/[1-(1-\delta)L] + u_t \tag{6.13}$$

Multiplying both sides of (6.13) by $[1-(1-\delta)L]$, we get

$$Y_t - (1-\delta)Y_{t-1} = \alpha[(1-(1-\delta)] + \beta\delta X_t + u_t - (1-\delta)u_{t-1} \tag{6.14}$$

(6.14) looks exactly like (6.10) with $\lambda = (1-\delta)$.

6.2.2 Partial Adjustment Model (PAM)

Under this model there is a cost of being out of equilibrium and a cost of adjusting to that equilibrium, i.e.,

$$Cost = a(Y_t - Y_t^*)^2 + b(Y_t - Y_{t-1})^2 \tag{6.15}$$

where Y_t^* is the target or equilibrium level for Y, whereas Y_t is the current level of Y. The first term of (6.15) gives a quadratic loss function proportional to the distance of Y_t from the equilibrium level Y_t^*. The second quadratic term represents the cost of adjustment. Minimizing this quadratic cost function with respect to Y, we get $Y_t = \gamma Y_t^* + (1-\gamma)Y_{t-1}$, where $\gamma = a/(a+b)$. Note that if the cost of adjustment was zero, then $b = 0$, $\gamma = 1$, and the target is reached immediately. However, there are costs of adjustment, especially in building the desired capital stock. Hence,

$$Y_t = \gamma Y_t^* + (1-\gamma)Y_{t-1} + u_t \tag{6.16}$$

where we made this relationship stochastic. If the true relationship is $Y_t^* = \alpha + \beta X_t$, then from (6.16)

$$Y_t = \gamma\alpha + \gamma\beta X_t + (1-\gamma)Y_{t-1} + u_t \tag{6.17}$$

and this looks like (6.10) with $\lambda = (1-\gamma)$, except for the error term, which is not necessarily MA(1) with the Moving Average parameter λ.

6.3 Estimation and Testing of Dynamic Models with Serial Correlation

Both the AEM and the PAM give equations resembling the autoregressive form of the infinite distributed lag. In all cases, we end up with a lagged dependent variable and an error term that is either Moving Average of order one as in (6.10), or just classical or autoregressive as in (6.17). In this section we study the testing and estimation of such *autoregressive* or *dynamic* models.

If there is a Y_{t-1} in the regression equation and the u_t's are classical disturbances, as may be the case in equation (6.17), then Y_{t-1} is said to be *contemporaneously uncorrelated* with the disturbance term u_t. In fact, the disturbances satisfy assumptions 1-4 of Chapter 3 and $E(Y_{t-1}u_t) = 0$ even though $E(Y_{t-1}u_{t-1}) \neq 0$. In other words, Y_{t-1} is not correlated with the current disturbance u_t but it is correlated with the lagged disturbance u_{t-1}. In this case, as long as the disturbances are not serially correlated, OLS will be biased, but remains consistent and asymptotically efficient. This case is unlikely with economic data given that most macro time-series variables are highly trended. More likely, the u_t's are serially correlated. In this case, OLS is biased and inconsistent. Intuitively, Y_t is related to u_t, so Y_{t-1} is related to u_{t-1}. If u_t and u_{t-1} are correlated, then Y_{t-1} and u_t are correlated. This means that one of the regressors, lagged Y, is correlated with u_t and we have the problem of endogeneity. Let us demonstrate what happens to OLS for the simple autoregressive model with no constant

$$Y_t = \beta Y_{t-1} + \nu_t \quad |\beta| < 1 \quad t = 1, 2, \ldots, T \tag{6.18}$$

with $\nu_t = \rho\nu_{t-1} + \epsilon_t$, $|\rho| < 1$ and $\epsilon_t \sim \text{IIN}(0, \sigma_\epsilon^2)$. One can show, see problem 3, that

$$\widehat{\beta}_{OLS} = \sum_{t=2}^{T} Y_t Y_{t-1} / \sum_{t=2}^{T} Y_{t-1}^2 = \beta + \sum_{t=2}^{T} Y_{t-1}\nu_t / \sum_{t=2}^{T} Y_{t-1}^2$$

with $\text{plim}(\widehat{\beta}_{OLS} - \beta) = \text{asymp. bias}(\widehat{\beta}_{OLS}) = \rho(1-\beta^2)/(1+\rho\beta)$. This asymptotic bias is positive if $\rho > 0$ and negative if $\rho < 0$. Also, this asymptotic bias can be large for small values of β and large values of ρ. For example, if $\rho = 0.9$ and $\beta = 0.2$, the asymptotic bias for β is 0.73. This is more than 3 times the value of β.

Also, $\widehat{\rho} = \sum_{t=2}^{T} \widehat{\nu}_t \widehat{\nu}_{t-1} / \sum_{t=2}^{T} \widehat{\nu}_{t-1}^2$ where $\widehat{\nu}_t = Y_t - \widehat{\beta}_{OLS} Y_{t-1}$ has

$$\text{plim}(\widehat{\rho} - \rho) = -\rho(1-\beta^2)/(1+\rho\beta) = -\text{ asymp.bias}(\widehat{\beta}_{OLS})$$

This means that if $\rho > 0$, then $\widehat{\rho}$ would be negatively biased. However, if $\rho < 0$, then $\widehat{\rho}$ is positively biased. In both cases, $\widehat{\rho}$ is biased towards zero. In fact, the asymptotic bias of the D.W. statistic is twice the asymptotic bias of $\widehat{\beta}_{OLS}$, see problem 3. This means that the D.W. statistic is biased towards not rejecting the null hypothesis of zero serial correlation. Therefore, if the D.W. statistic rejects the null of $\rho = 0$, it is doing that when the odds are against it, and therefore confirming our rejection of the null and the presence of serial correlation. If on the other hand it does not reject the null, then the D.W. statistic is uninformative and has to be replaced by another conclusive test for serial correlation. Such an alternative test in the presence of a lagged dependent variable has been developed by Durbin (1970), and the statistic computed is called Durbin's h. Using (6.10) or (6.17), one computes OLS ignoring its possible bias and $\widehat{\rho}$ from OLS residuals as shown above. Durbin's h is given by

$$h = \widehat{\rho}[n/(1 - n\ \widehat{\text{var}}(\text{coeff. of } Y_{t-1}))]^{1/2}. \tag{6.19}$$

This is asymptotically distributed $N(0,1)$ under null hypothesis of $\rho = 0$. If $n[\widehat{\text{var}}(\text{coeff. of } Y_{t-1})]$ is greater than one, then h cannot be computed, and Durbin suggests running the OLS residuals e_t on e_{t-1} and the regressors in the model (including the lagged dependent variable), and testing whether the coefficient of e_{t-1} in this regression is significant. In fact, this test can be generalized to higher order autoregressive errors. Let u_t follow an AR(p) process

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + .. + \rho_p u_{t-p} + \epsilon_t$$

then this test involves running e_t on $e_{t-1}, e_{t-2}, \ldots, e_{t-p}$ and the regressors in the model including Y_{t-1}. The test statistic for $H_0; \rho_1 = \rho_2 = .. = \rho_p = 0$; is TR^2 which is distributed χ_p^2. This is the Lagrange multiplier test developed independently by Breusch (1978) and Godfrey (1978) and discussed in Chapter 5. In fact, this test has other useful properties. For example, this test is the same whether the null imposes an AR(p) model or an MA(p) model on the disturbances, see Chapter 14. Kiviet (1986) argues that even though these are large sample tests, the Breusch-Godfrey test is preferable to Durbin's h in small samples.

6.3.1 A Lagged Dependent Variable Model with AR(1) Disturbances

A model with a lagged dependent variable and an autoregressive error term is estimated using instrumental variables (IV). This method will be studied extensively in Chapter 11. In short, the IV method corrects for the correlation between Y_{t-1} and the error term by replacing Y_{t-1} with its predicted value $\widehat{Y}_{t-1}$. The latter is obtained by regressing Y_{t-1} on some exogenous variables,

say a set of Z's, which are called a set of instruments for Y_{t-1}. Since these variables are exogenous and uncorrelated with u_t, $\widehat{Y}_{t-1}$ will not be correlated with u_t. Suppose the regression equation is

$$Y_t = \alpha + \beta Y_{t-1} + \gamma X_t + u_t \qquad t = 2, \ldots, T \tag{6.20}$$

and that at least one exogenous variable Z_t exists which will be our instrument for Y_{t-1}. Regressing Y_{t-1} on X_t, Z_t and a constant, we get

$$Y_{t-1} = \widehat{Y}_{t-1} + \widehat{\nu}_t = \widehat{a}_1 + \widehat{a}_2 Z_t + \widehat{a}_3 X_t + \widehat{\nu}_t. \tag{6.21}$$

Then $\widehat{Y}_{t-1} = \widehat{a}_1 + \widehat{a}_2 Z_t + \widehat{a}_3 X_t$ and is independent of u_t, because it is a linear combination of exogenous variables. But, Y_{t-1} is correlated with u_t. This means that $\widehat{\nu}_t$ is the part of Y_{t-1} that is correlated with u_t. Substituting $Y_{t-1} = \widehat{Y}_{t-1} + \widehat{\nu}_t$ in (6.20) we get

$$Y_t = \alpha + \beta \widehat{Y}_{t-1} + \gamma X_t + (u_t + \beta \widehat{\nu}_t) \tag{6.22}$$

$\widehat{Y}_{t-1}$ is uncorrelated with the new error term $(u_t + \beta\widehat{\nu}_t)$ because $\Sigma \widehat{Y}_{t-1}\widehat{\nu}_t = 0$ from (6.21). Also, X_t is uncorrelated with u_t by assumption. But, from (6.21), X_t also satisfies $\Sigma X_t \widehat{\nu}_t = 0$. Hence, X_t is uncorrelated with the new error term $(u_t + \beta\widehat{\nu}_t)$. This means that OLS applied to (6.22) will lead to consistent estimates of α, β and γ. The only remaining question is where do we find instruments like Z_t? This Z_t should be (i) uncorrelated with u_t, (ii) preferably predicting Y_{t-1} fairly well, but, not predicting it perfectly, otherwise $\widehat{Y}_{t-1} = Y_{t-1}$. If this happens, we are back to OLS which we know is inconsistent, (iii) $\Sigma z_t^2/T$ should be finite and different from zero. Recall that $z_t = Z_t - \bar{Z}$. In this case, X_{t-1} seems like a natural instrumental variable candidate. It is an exogenous variable which would very likely predict Y_{t-1} fairly well, and satisfies $\Sigma x_{t-1}^2/T$ being finite and different from zero. In other words, (6.21) regresses Y_{t-1} on a constant, X_{t-1} and X_t, and gets $\widehat{Y}_{t-1}$. Additional lags on X_t can be used as instruments to improve the small sample properties of this estimator. Substituting $\widehat{Y}_{t-1}$ in equation (6.22) results in consistent estimates of the regression parameters. Wallis (1967) substituted these consistent estimates in the original equation (6.20) and obtained the residuals $\widetilde{u}_t$. Then he computed

$$\widehat{\rho} = [\textstyle\sum_{t=2}^{T} \widetilde{u}_t \widetilde{u}_{t-1}/(T-1)]/[\sum_{t=1}^{T} \widetilde{u}_t^2/T] + (3/T)$$

where the last term corrects for the bias in $\widehat{\rho}$. At this stage, one can perform a Prais-Winsten procedure on (6.20) using $\widehat{\rho}$ instead of ρ, see Fomby and Guilkey (1983).

An alternative two-step procedure has been proposed by Hatanaka (1974). After estimating (6.22) and obtaining the residuals $\widetilde{u}_t$ from (6.20), Hatanaka (1974) suggests running $Y_t^* = Y_t - \widetilde{\rho} Y_{t-1}$ on $Y_{t-1}^* = Y_{t-1} - \widetilde{\rho} Y_{t-2}$, $X_t^* = X_t - \widetilde{\rho} X_{t-1}$ and $\widetilde{u}_{t-1}$. Note that this is the Cochrane-Orcutt transformation which ignores the first observation. Also, $\widetilde{\rho} = \sum_{t=3}^{T} \widetilde{u}_t \widetilde{u}_{t-1} / \sum_{t=3}^{T} \widetilde{u}_t^2$ ignores the small sample bias correction factor suggested by Wallis (1967). Let $\widetilde{\delta}$ be the coefficient of $\widetilde{u}_{t-1}$, then the efficient estimator of ρ is given by $\widetilde{\widetilde{\rho}} = \widetilde{\rho} + \widetilde{\delta}$. Hatanaka shows that the resulting estimators are asymptotically equivalent to the MLE in the presence of Normality.

Empirical Example: Consider the Consumption-Income data from the Economic Report of the President over the period 1950-1993 given in Table 5.1. Problem 5 asks the reader to verify that Durbin's h obtained from the lagged dependent variable model described in (6.20) yields a value of 3.367. This is asymptotically distributed as $N(0, 1)$ under the null hypothesis of no

serial correlation of the disturbances. This null is soundly rejected. The Bruesch and Godfrey test runs the regression of OLS residuals on their lagged values and the regressors in the model. This yields a $TR^2 = 7.972$. This is distributed as χ^2_1 under the null. Therefore, we reject the hypothesis of no first-order serial correlation. Next, we estimate (6.20) using current and lagged values of income (Y_t, Y_{t-1} and Y_{t-2}) as a set of instruments for lagged consumption (C_{t-1}). The regression given by (6.22), yields:

$$\begin{array}{lllll} C_t = & -0.053 + & 0.196 & \widehat{C}_{t-1} + & 0.802 \quad Y_t + \text{residuals} \\ & (0.0799) & (0.1380) & & (0.1386) \end{array}$$

Substituting these estimates in (6.20), one gets the residuals $\widetilde{u}_t$. Based on these $\widetilde{u}_t$'s, the Wallis (1967) estimate of ρ yields $\widehat{\rho} = 0.681$. Using this $\widetilde{\rho}$, the Prais-Winsten regression on (6.20) gives the following result:

$$\begin{array}{llll} C_t = 0.0007 + & 0.169 \quad C_{t-1} + & 0.822 \quad Y_t + \text{residuals} \\ (0.007) & (0.088) & (0.088) \end{array}$$

Alternatively, based on $\widetilde{u}_t$, one can compute Hatanaka's (1974) estimate of ρ given by $\widetilde{\rho} = 0.597$ and run Hatanaka's regression

$$\begin{array}{llllll} C^*_t = -0.036 + & 0.182 \quad C^*_{t-1} + & 0.820 \quad Y^*_t + & 0.068 \quad \widetilde{u}_{t-1} + \text{residuals} \\ (0.054) & (0.098) & (0.099) & (0.142) \end{array}$$

where $C^*_t = C_t - \widetilde{\rho} C_{t-1}$. The efficient estimate of ρ is given by $\widetilde{\widetilde{\rho}} = \widetilde{\rho} + 0.068 = 0.665$.

6.3.2 A Lagged Dependent Variable Model with MA(1) Disturbances

Zellner and Geisel (1970) estimated the Koyck autoregressive representation of the infinite distributed lag, given in (6.10). In fact, we saw that this could also arise from the AEM, see (6.14). In particular, it is a regression with a lagged dependent variable and an MA(1) error term with the added restriction that the coefficient of Y_{t-1} is the same as the MA(1) parameter. For simplicity, we write

$$Y_t = \alpha + \lambda Y_{t-1} + \beta X_t + (u_t - \lambda u_{t-1}) \tag{6.23}$$

Let $w_t = Y_t - u_t$, then (6.23) becomes

$$w_t = \alpha + \lambda w_{t-1} + \beta X_t \tag{6.24}$$

By continuous substitution of lagged values of w_t in (6.24) we get

$$w_t = \alpha(1 + \lambda + \lambda^2 + .. + \lambda^{t-1}) + \lambda^t w_0 + \beta(X_t + \lambda X_{t-1} + .. + \lambda^{t-1} X_1)$$

and replacing w_t by $(Y_t - u_t)$, we get

$$Y_t = \alpha(1 + \lambda + \lambda^2 + .. + \lambda^{t-1}) + \lambda^t w_0 + \beta(X_t + \lambda X_{t-1} + .. + \lambda^{t-1} X_1) + u_t \tag{6.25}$$

knowing λ, this equation can be estimated via OLS assuming that the disturbances u_t are not serially correlated. Since λ is not known, Zellner and Geisel (1970) suggest a search procedure over λ, where $0 < \lambda < 1$. The regression with the minimum residual sums of squares gives the

optimal λ, and the corresponding regression gives the estimates of α, β and w_0. The last coefficient $w_o = Y_o - u_o = E(Y_o)$ can be interpreted as the expected value of the initial observation on the dependent variable. Klein (1958) considered the direct estimation of the infinite Koyck lag, given in (6.8) and arrived at (6.25). The search over λ results in MLEs of the coefficients. Note, however, that the estimate of w_o is not consistent. Intuitively, as t tends to infinity, λ^t tends to zero implying no new information to estimate w_o. In fact, some applied researchers ignore the variable λ^t in the regression given in (6.25). This practice, known as truncating the remainder, is not recommended since the Monte Carlo experiments of Maddala and Rao (1971) and Schmidt (1975) have shown that even for $T = 60$ or 100, it is not desirable to omit λ^t from (6.25).

In summary, we have learned how to estimate a dynamic model with a lagged dependent variable and serially correlated errors. In case the error is autoregressive of order one, we have outlined the steps to implement the Wallis Two-Stage estimator and Hatanaka's two-step procedure. In case the error is Moving Average of order one, we have outlined the steps to implement the Zellner-Geisel procedure.

6.4 Autoregressive Distributed Lag

So far, section 6.1 considered finite distributed lags on the explanatory variables, whereas section 6.2 considered an autoregressive relation including the first lag of the dependent variable and current values of the explanatory variables. In general, economic relationships may be generated by an *Autoregressive Distributed Lag* (ADL) scheme. The simplest form is the ADL (1,1) model which is given by

$$Y_t = \alpha + \lambda Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + u_t \tag{6.26}$$

where both Y_t and X_t are lagged once. By specifying higher order lags for Y_t and X_t, say an ADL (p, q) with p lags on Y_t and q lags on X_t, one can test whether the specification now is general enough to ensure White noise disturbances. Next, one can test whether some restrictions can be imposed on this general model, like reducing the order of the lags to arrive at a simpler ADL model, or estimating the simpler static model with the Cochrane-Orcutt correction for serial correlation, see problem 20 in Chapter 7. This general to specific modelling strategy is prescribed by David Hendry and is utilized by the econometric software PC-Give, see Gilbert (1986).

Returning to the ADL $(1, 1)$ model in (6.26) one can invert the autoregressive form as follows:

$$Y_t = \alpha(1 + \lambda + \lambda^2 + ..) + (1 + \lambda L + \lambda^2 L^2 + ..)(\beta_0 X_t + \beta_1 X_{t-1} + u_t) \tag{6.27}$$

provided $|\lambda| < 1$. This equation gives the effect of a unit change in X_t on future values of Y_t. In fact, $\partial Y_t / \partial X_t = \beta_0$ while $\partial Y_{t+1} / \partial X_t = \beta_1 + \lambda\beta_0$, etc. This gives the immediate short-run responses with the long-run effect being the sum of all these partial derivatives yielding $(\beta_0 + \beta_1)/(1 - \lambda)$. This can be alternatively derived from (6.26) at the long-run static equilibrium (Y^*, X^*) where $Y_t = Y_{t-1} = Y^*$, $X_t = X_{t-1} = X^*$ and the disturbance is set equal to zero, i.e.,

$$Y^* = \frac{\alpha}{1 - \lambda} + \frac{\beta_0 + \beta_1}{1 - \lambda} X^* \tag{6.28}$$

Replacing Y_t by $Y_{t-1} + \Delta Y_t$ and X_t by $X_{t-1} + \Delta X_t$ in (6.26) one gets

$$\Delta Y_t = \alpha + \beta_0 \Delta X_t - (1 - \lambda) Y_{t-1} + (\beta_0 + \beta_1) X_{t-1} + u_t$$

This can be rewritten as

$$\Delta Y_t = \beta_0 \Delta X_t - (1-\lambda)\left[Y_{t-1} - \frac{\alpha}{1-\lambda} - \frac{\beta_0+\beta_1}{1-\lambda} X_{t-1}\right] + u_t \tag{6.29}$$

Note that the term in brackets contains the long-run equilibrium parameters derived in (6.28). In fact, the term in brackets represents the deviation of Y_{t-1} from the long-run equilibrium term corresponding to X_{t-1}. Equation (6.29) is known as the *Error Correction Model* (ECM), see Davidson, Hendry, Srba and Yeo (1978). Y_t is obtained from Y_{t-1} by adding the short-run effect of the change in X_t and a long-run equilibrium adjustment term. Since, the disturbances are White noise, this model is estimated by OLS.

Note

1. Other distributions besides the geometric distribution can be considered. In fact, a Pascal distribution was considered by Solow (1960), a rational-lag distribution was considered by Jorgenson (1966), and a Gamma distribution was considered by Schmidt (1974, 1975). See Maddala (1977) for an excellent review.

Problems

1. Consider the Consumption-Income data given in Table 5.1 and provided on the Springer web site as CONSUMP.DAT. Estimate a Consumption-Income regression in logs that allows for a six year lag on income as follows:

   (a) Use the linear arithmetic lag given in equation (6.2). Show that this result can also be obtained as an Almon lag first-degree polynomial with a far end point constraint.

   (b) Use an Almon lag second-degree polynomial, described in equation (6.4), imposing the near end point constraint.

   (c) Use an Almon lag second-degree polynomial imposing the far end point constraint.

   (d) Use an Almon lag second-degree polynomial imposing both end point constraints.

   (e) Using Chow's F-statistic, test the arithmetic lag restrictions given in part (a).

   (f) Using Chow's F-statistic, test the Almon lag restrictions implied by the model in part (b).

   (g) Repeat part (f) for the restrictions imposed in parts (c) and (d).

2. Consider fitting an Almon lag third degree polynomial $\beta_i = a_0 + a_1 i + a_2 i^2 + a_3 i^3$ for $i = 0, 1, \ldots, 5$, on the Consumption-Income relationship in logarithms. In this case, there are five lags on income, i.e., $s = 5$.

   (a) Set up the estimating equation for the a_i's and report the estimates using OLS.

   (b) What is your estimate of β_3? What is the standard error? Can you relate the var$(\widehat{\beta}_3)$ to the variances and covariances of the a_i's?

   (c) How would the OLS regression in part (a) change if we impose the near end point constraint $\beta_{-1} = 0$?

   (d) Test the near end point constraint.

(e) Test the Almon lag specification given in part (a) against an unrestricted five year lag specification on income.

3. For the simple dynamic model with AR(1) disturbances given in (6.18),

(a) Verify that $\text{plim}(\widehat{\beta}_{OLS}-\beta) = \rho(1-\beta^2)/(1+\rho\beta)$. **Hint:** From (6.18), $Y_{t-1} = \beta Y_{t-2}+\nu_{t-1}$ and $\rho Y_{t-1} = \rho\beta Y_{t-2} + \rho\nu_{t-1}$. Subtracting this last equation from (6.18) and re-arranging terms, one gets $Y_t = (\beta+\rho)Y_{t-1} - \rho\beta Y_{t-2}+\epsilon_t$. Multiply both sides by Y_{t-1} and sum $\sum_{t=2}^{T} Y_t Y_{t-1} = (\beta+\rho)\sum_{t=2}^{T} Y_{t-1}^2 - \rho\beta\sum_{t=2}^{T} Y_{t-1}Y_{t-2} + \sum_{t=2}^{T} Y_{t-1}\epsilon_t$. Now divide by $\sum_{t=2}^{T} Y_{t-1}^2$ and take probability limits. See Griliches (1961).

(b) For various values of $|\rho| < 1$ and $|\beta| < 1$, tabulate the asymptotic bias computed in part (a).

(c) Verify that $\text{plim}(\widehat{\rho}-\rho) = -\rho(1-\beta^2)/(1+\rho\beta) = -\text{plim}(\widehat{\beta}_{OLS}-\beta)$.

(d) Using part (c), show that plim $d = 2(1-\text{ plim }\widehat{\rho}) = 2[1 - \dfrac{\beta\rho(\beta+\rho)}{1+\beta\rho}]$ where $d = \sum_{t=2}^{T}(\widehat{\nu}_t - \widehat{\nu}_{t-1})^2/\sum_{t=1}^{T}\widehat{\nu}_t^2$ denotes the Durbin-Watson statistic.

(e) Knowing the true disturbances, the Durbin-Watson statistic would be $d^* = \sum_{t=2}^{T}(\nu_t - \nu_{t-1})^2/\sum_{t=1}^{T}\nu_t^2$ and its plim $d^* = 2(1-\rho)$. Using part (d), show that plim $(d-d^*) = \dfrac{2\rho(1-\beta^2)}{1+\beta\rho} = 2\text{plim}(\widehat{\beta}_{OLS}-\beta)$ obtained in part (a). See Nerlove and Wallis (1966). For various values of $|\rho| < 1$ and $|\beta| < 1$, tabulate d^* and d and the asymptotic bias in part (d).

4. For the simple dynamic model given in (6.18), let the disturbances follow an MA(1) process $\nu_t = \epsilon_t + \theta\epsilon_{t-1}$ with $\epsilon_t \sim \text{IIN}(0,\sigma_\epsilon^2)$.

(a) Show that $\text{plim}(\widehat{\beta}_{OLS}-\beta) = \dfrac{\delta(1-\beta^2)}{1+2\beta\delta}$ where $\delta = \theta/(1+\theta^2)$.

(b) Tabulate this asymptotic bias for various values of $|\beta| < 1$ and $0 < \theta < 1$.

(c) Show that $\text{plim}(\dfrac{1}{T}\sum_{t=2}^{T}\widehat{\nu}_t^2) = \sigma_\epsilon^2[1+\theta(\theta-\theta^*)]$ where $\theta^* = \delta(1-\beta^2)/(1+2\beta\delta)$ and $\widehat{\nu}_t = Y_t - \widehat{\beta}_{OLS}Y_{t-1}$.

5. Consider the lagged dependent variable model given in (6.20). Using the Consumption-Income data from the Economic Report of the President over the period 1950-1993 which is given in Table 5.1.

(a) Test for first-order serial correlation in the disturbances using Durbin's h given in (6.19).

(b) Test for first-order serial correlation in the disturbances using the Breusch (1978) and Godfrey (1978) test.

(c) Test for second-order serial correlation in the disturbances.

6. Using the U.S. gasoline data in Chapter 4, problem 15 given in Table 4.2 and obtained from the USGAS.ASC file, estimate the following two models:

$$\text{Static: } \log\left(\frac{QMG}{CAR}\right)_t = \gamma_1 + \gamma_2\log\left(\frac{RGNP}{POP}\right)_t + \gamma_3\log\left(\frac{CAR}{POP}\right)_t + \gamma_4\log\left(\frac{PMG}{PGNP}\right)_t + \epsilon_t$$

$$\text{Dynamic: } \log\left(\frac{QMG}{CAR}\right)_t = \gamma_1 + \gamma_2\log\left(\frac{RGNP}{POP}\right)_t + \gamma_3\left(\frac{CAR}{POP}\right)_t + \gamma_4\log\left(\frac{PMG}{PGNP}\right)_t + \lambda\log\left(\frac{QMG}{CAR}\right)_{t-1} + \epsilon_t$$

(a) Compare the implied short-run and long-run elasticities for price (PMG) and income ($RGNP$).

(b) Compute the elasticities after $3, 5$ and 7 years. Do these lags seem plausible?

(c) Can you apply the Durbin-Watson test for serial correlation to the dynamic version of this model? Perform Durbin's h-test for the dynamic gasoline model. Also, the Breusch-Godfrey test for first-order serial correlation.

7. Using the U.S. gasoline data in Chapter 4, problem 15, given in Table 4.2 estimate the following model with a six year lag on prices:

$$\log\left(\frac{QMG}{CAR}\right)_t = \gamma_1 + \gamma_2 \log\left(\frac{RGNP}{POP}\right)_t + \gamma_3 \log\left(\frac{CAR}{POP}\right)_t + \gamma_4 \sum_{i=0}^{6} w_i \log\left(\frac{PMG}{PGNP}\right)_{t-i}$$

(a) Report the unrestricted OLS estimates.

(b) Now, estimate a second degree polynomial lag for the same model. Compare the results with part (a) and explain why you got such different results.

(c) Re-estimate part (b) comparing the six year lag to a four year, and eight year lag. Which one would you pick?

(d) For the six year lag model, does a third degree polynomial give a better fit?

(e) For the model outlined in part (b), reestimate with a far end point constraint. Now, reestimate with only a near end point constraint. Are such restrictions justified in this case?

References

This chapter is based on the material in Maddala (1977), Johnston (1984), Kelejian and Oates (1989) and Davidson and MacKinnon (1993). Additional references on the material in this chapter include:

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in B. Petrov and F. Csake, eds. *2nd. International Symposium on Information Theory*, Budapest: Akademiai Kiado.

Almon, S. (1965), "The Distributed Lag Between Capital Appropriations and Net Expenditures," *Econometrica*, 30: 407-423.

Breusch, T.S. (1978), "Testing for Autocorrelation in Dynamic Linear Models," *Australian Economic Papers*, 17: 334-355.

Davidson, J.E.H., D.F. Hendry, F. Srba and S. Yeo (1978), "Econometric Modelling of the Aggregate Time-Series Relationship Between Consumers' Expenditure and Income in the United Kingdom," *Economic Journal*, 88: 661-692.

Dhrymes, P.J. (1971), *Distributed Lags: Problems of Estimation and Formulation* (Holden-Day: San Francisco).

Durbin, J. (1970), "Testing for Serial Correlation in Least Squares Regression when Some of the Regressors are Lagged Dependent Variables," *Econometrica*, 38: 410-421.

Fomby, T.B. and D.K. Guilkey (1983), "An Examination of Two-Step Estimators for Models with Lagged Dependent and Autocorrelated Errors," *Journal of Econometrics*, 22: 291-300.

Gilbert, C.L. (1986), "Professor Hendry's Econometric Methodology," *Oxford Bulletin of Economics and Statistics*, 48: 283-307.

Godfrey, L.G. (1978), "Testing Against General Autoregressive and Moving Average Error Models when the Regressors Include Lagged Dependent Variables," *Econometrica*, 46: 1293-1302.

Griliches, Z. (1961), "A Note on Serial Correlation Bias in Estimates of Distributed Lags," *Econometrica*, 29: 65-73.

Hatanaka, M. (1974), "An Efficient Two-Step Estimator for the Dynamic Adjustment Model with Autocorrelated Errors," *Journal of Econometrics*, 2: 199-220.

Jorgenson, D.W. (1966), "Rational Distributed Lag Functions," *Econometrica*, 34: 135-149.

Kiviet, J.F. (1986), "On The Vigor of Some Misspecification Tests for Modelling Dynamic Relationships," *Review of Economic Studies*, 53: 241-262.

Klein, L.R. (1958), "The Estimation of Distributed Lags," *Econometrica*, 26: 553-565.

Koyck, L.M. (1954), *Distributed Lags and Investment Analysis* (North-Holland: Amsterdam).

Maddala, G.S. and A.S. Rao (1971), "Maximum Likelihood Estimation of Solow's and Jorgenson's Distributed Lag Models," *Review of Economics and Statistics*, 53: 80-88.

Nerlove, M. and K.F. Wallis (1967), "Use of the Durbin-Watson Statistic in Inappropriate Situations," *Econometrica*, 34: 235-238.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6: 461-464.

Schmidt, P. (1974), "An Argument for the Usefulness of the Gamma Distributed Lag Model," *International Economic Review*, 15: 246-250.

Schmidt, P. (1975), "The Small Sample Effects of Various Treatments of Truncation Remainders on the Estimation of Distributed Lag Models," *Review of Economics and Statistics*, 57: 387-389.

Schmidt, P. and R. N. Waud (1973), "The Almon lag Technique and the Monetary versus Fiscal Policy Debate," *Journal of the American Statistical Association*, 68: 11-19.

Solow, R.M. (1960), "On a Family of Lag Distributions," *Econometrica*, 28: 393-406.

Wallace, T.D. (1972), "Weaker Criteria and Tests for Linear Restrictions in Regression," *Econometrica*, 40: 689-698.

Wallis, K.F. (1967), "Lagged Dependent Variables and Serially Correlated Errors: A Reappraisal of Three-Pass Least Squares, " *Review of Economics and Statistics*, 49: 555-567.

Zellner, A. and M.Geisel (1970), "Analysis of Distributed Lag Models with Application to Consumption Function Estimation," *Econometrica*, 38: 865-888.

Part II

CHAPTER 7
The General Linear Model: The Basics

7.1 Introduction

Consider the following regression equation

$$y = X\beta + u \tag{7.1}$$

where

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ : \\ Y_n \end{bmatrix}; X = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ : & : & : & : \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix}; \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ : \\ \beta_k \end{bmatrix}; u = \begin{bmatrix} u_1 \\ u_2 \\ : \\ u_n \end{bmatrix}$$

with n denoting the number of observations and k the number of variables in the regression, with $n > k$. In this case, y is a column vector of dimension $(n \times 1)$ and X is a matrix of dimension $(n \times k)$. Each column of X denotes a variable and each row of X denotes an observation on these variables. If y is log(wage) as in the empirical example in Chapter 4, see Table 4.1 then the columns of X contain a column of ones for the constant (usually the first column), weeks worked, years of full time experience, years of education, sex, race, marital status, etc.

7.2 Least Squares Estimation

Least squares minimizes the residual sum of squares where the residuals are given by $e = y - X\widehat{\beta}$ and $\widehat{\beta}$ denotes a guess on the regression parameters β. The residual sum of squares

$$RSS = \textstyle\sum_{i=1}^{n} e_i^2 = e'e = (y - X\beta)'(y - X\beta) = y'y - y'X\beta - \beta'X'y + \beta'X'X\beta$$

The last four terms are scalars as can be verified by their dimensions. It is essential that the reader keep track of the dimensions of the matrices used. This will insure proper multiplication, addition, subtraction of matrices and help the reader obtain the right answers. In fact the middle two terms are the same because the transpose of a scalar is a scalar. For a quick review of some matrix properties, see the Appendix to this chapter. Differentiating the RSS with respect to β one gets

$$\partial RSS/\partial\beta = -2X'y + 2X'X\beta \tag{7.2}$$

where use is made of the following two rules of differentiating matrices. The first is that $\partial a'b/\partial b = a$ and the second is

$$\partial(b'Ab)/\partial b = (A + A')b = 2Ab$$

where the last equality holds if A is a symmetric matrix. In the RSS equation a is $y'X$ and A is $X'X$. The first-order condition for minimization equates the expression in (7.2) to zero. This yields

$$X'X\beta = X'y \tag{7.3}$$

which is known as the OLS normal equations. As long as X is of full column rank , i.e., of rank k, then $X'X$ is nonsingular and the solution to the above equations is $\widehat{\beta}_{OLS} = (X'X)^{-1}X'y$. Full column rank means that no column of X is a perfect linear combination of the other columns. In other words, no variable in the regression can be obtained from a linear combination of the other variables. Otherwise, at least one of the OLS normal equations becomes redundant. This means that we only have $(k-1)$ linearly independent equations to solve for k unknown β's. This yields no solution for $\widehat{\beta}_{OLS}$ and we say that $X'X$ is singular. $X'X$ is the sum of squares cross product matrix (SSCP). If it has a column of ones then it will contain the sums, the sum of squares, and the cross-product sum between any two variables

$$X'X = \begin{bmatrix} n & \sum_{i=1}^n X_{i2} & \dots & \sum_{i=1}^n X_{ik} \\ \sum_{i=1}^n X_{i2} & \sum_{i=1}^n X_{i2}^2 & \dots & \sum_{i=1}^n X_{i2}X_{ik} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^n X_{ik} & \sum_{i=1}^n X_{ik}X_{i2} & \dots & \sum_{i=1}^n X_{ik}^2 \end{bmatrix}$$

Of course y could be added to this matrix as another variable which will generate $X'y$ and $y'y$ automatically for us, i.e., the column pertaining to the variable y will generate $\sum_{i=1}^n y_i$, $\sum_{i=1}^n X_{i1}y_i, \dots, \sum_{i=1}^n X_{ik}y_i$, and $\sum_{i=1}^n y_i^2$. To see this, let

$$Z = [y, X] \quad \text{then} \quad Z'Z = \begin{bmatrix} y'y & y'X \\ X'y & X'X \end{bmatrix}$$

This matrix summarizes the data and we can compute any regression of one variable in Z on any subset of the remaining variables in Z using only $Z'Z$. Denoting the least squares residuals by $e = y - X\widehat{\beta}_{OLS}$, the OLS normal equations given in (7.3) can be written as

$$X'(y - X\widehat{\beta}_{OLS}) = X'e = 0 \tag{7.4}$$

Note that if the regression includes a constant, the first column of X will be a vector of ones and the first equation of (7.4) becomes $\sum_{i=1}^n e_i = 0$. This proves the well known result that if there is a constant in the regression, the OLS residuals sum to zero. Equation (7.4) also indicates that the regressor matrix X is orthogonal to the residuals vector e. This will become clear when we define e in terms of the orthogonal projection matrix on X. This representation allows another interpretation of OLS as a method of moments estimator which was considered in Chapter 2. This follows from the classical assumptions where X satisfies $E(X'u) = 0$. The sample counterpart of this condition yields $X'e/n = 0$. These are the OLS normal equations and therefore, yield the OLS estimates without minimizing the residual sums of squares. See Kelejian and Oates (1989) and the discussion of instrumental variable estimation in Chapter 11.

Since data in economics are not generated using experiments like the physical sciences, the X's are stochastic and we only observe one realization of this data. Consider for example, annual observations for GNP, money supply, unemployment rate, etc. One cannot repeat draws for this data in the real world or fix the X's to generate new y's (unless one is performing a Monte Carlo study). So we have to condition on the set of X's observed, see Chapter 5.

Classical Assumptions: $u \sim (0, \sigma^2 I_n)$ which means that (i) each disturbance u_i has zero mean, (ii) constant variance, and (iii) u_i and u_j for $i \neq j$ are not correlated. The u's are known as spherical disturbances. Also, (iv) the conditional expectation of u given X is zero, $E(u/X) = 0$. Note that the conditioning here is with respect to *every* regressor in X and for *all* observations

$i = 1, 2, \ldots n$. In other words, it is conditional on all the elements of the matrix X. Using (7.1), this implies that $E(y/X) = X\beta$ is *linear* in β, $\text{var}(u_i/X) = \sigma^2$ and $\text{cov}(u_i, u_j/X) = 0$. Additionally, we assume that plim $X'X/n$ is finite and positive definite and plim $X'u/n = 0$ as $n \to \infty$.

Given these classical assumptions, and conditioning on the X's observed, it is easy to show that $\widehat{\beta}_{OLS}$ is unbiased for β. In fact using (7.1) one can write

$$\widehat{\beta}_{OLS} = \beta + (X'X)^{-1}X'u \tag{7.5}$$

Taking expectations, conditioning on the X's, and using assumptions (i) and (iv), one attains the unbiasedness result. Furthermore, one can derive the variance-covariance matrix of $\widehat{\beta}_{OLS}$ from (7.5) since

$$\text{var}(\widehat{\beta}_{OLS}) = E(\widehat{\beta}_{OLS} - \beta)(\widehat{\beta}_{OLS} - \beta)' = E(X'X)^{-1}X'uu'X(X'X)^{-1} = \sigma^2(X'X)^{-1} \tag{7.6}$$

this uses assumption (iv) along with the fact that $E(uu') = \sigma^2 I_n$. This variance-covariance matrix is $(k \times k)$ and gives the variances of the $\widehat{\beta}_i$'s across the diagonal and the pairwise covariances of say $\widehat{\beta}_i$ and $\widehat{\beta}_j$ off the diagonal. The next theorem shows that among all linear unbiased estimators of $c'\beta$, it is $c'\widehat{\beta}_{OLS}$ which has the smallest variance. This is known as the Gauss-Markov Theorem.

Theorem 1: Consider the linear estimator $a'y$ for $c'\beta$, where both a and c are arbitrary vectors of constants. If $a'y$ is unbiased for $c'\beta$ then $\text{var}(a'y) \geq \text{var}(c'\widehat{\beta}_{OLS})$.

Proof: For $a'y$ to be unbiased for $c'\beta$ it must follow from (7.1) that $E(a'y) = a'X\beta + E(a'u) = a'X\beta = c'\beta$ which means that $a'X = c'$. Also, $\text{var}(a'y) = E(a'y - c'\beta)(a'y - c'\beta)' = E(a'uu'a) = \sigma^2 a'a$. Comparing this variance with that of $c'\widehat{\beta}_{OLS}$, one gets $\text{var}(a'y) - \text{var}(c'\widehat{\beta}_{OLS}) = \sigma^2 a'a - \sigma^2 c'(X'X)^{-1}c$. But, $c' = a'X$, therefore this difference becomes $\sigma^2[a'a - a'P_X a] = \sigma^2 a'\bar{P}_X a$ where P_X is a projection matrix on the X-plane defined as $X(X'X)^{-1}X'$ and $\bar{P}_X$ is defined as $I_n - P_X$. In fact, $P_X y = X\widehat{\beta}_{OLS} = \widehat{y}$ and $\bar{P}_X y = y - P_X y = y - \widehat{y} = e$. So that $\widehat{y}$ projects the vector y on the X-plane and e is the projection of y on the plane orthogonal to X or perpendicular to X, see Figure 7.1. Both P_X and $\bar{P}_X$ are idempotent which means that the above difference $\sigma^2 a\bar{P}_X a$ is greater or equal to zero since $\bar{P}_X$ is positive semi-definite. To see this, define $z = \bar{P}_X a$, then the above difference is equal to $\sigma^2 z'z \geq 0$.

The implications of the theorem are important. It means for example, that for the choice of $c' = (1, 0, \ldots, 0)$ one can pick $\beta_1 = c'\beta$ for which the best linear unbiased estimator would be $\widehat{\beta}_{1,OLS} = c'\widehat{\beta}_{OLS}$. Similarly any β_j can be chosen by using $c' = (0, \ldots, 1, \ldots, 0)$ which has 1 in the j-th position and zero elsewhere. Again, the BLUE of $\beta_j = c'\beta$ is $\widehat{\beta}_{j,OLS} = c'\widehat{\beta}_{OLS}$. Furthermore, any linear combination of these β's such as their sum $\sum_{j=1}^{k} \beta_j$ which corresponds to $c' = (1, 1, \ldots, 1)$ has the sum $\sum_{j=1}^{k} \widehat{\beta}_{j,OLS}$ as its BLUE.

The disturbance variance σ^2 is unknown and has to be estimated. Note that $E(u'u) = E(\text{tr}(uu')) = \text{tr}(E(uu')) = \text{tr}(\sigma^2 I_n) = n\sigma^2$, so that $u'u/n$ seems like a natural unbiased estimator for σ^2. However, u is not observed and is estimated by the OLS residuals e. It is therefore, natural to investigate $E(e'e)$. In what follows, we show that $s^2 = e'e/(n-k)$ is an unbiased estimator for σ^2. To prove this, we need the fact that

$$e = y - X\widehat{\beta}_{OLS} = y - X(X'X)^{-1}X'y = \bar{P}_X y = \bar{P}_X u \tag{7.7}$$

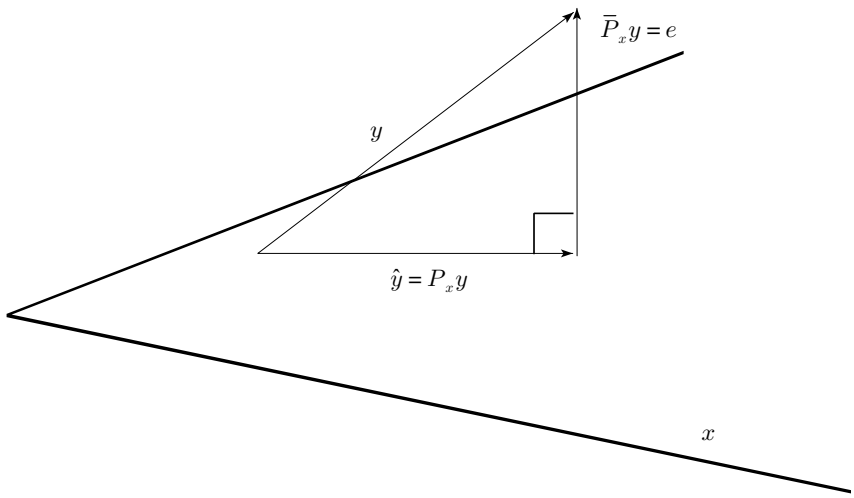


Figure 7.1 The Orthogonal Decomposition of y

where the last equality follows from the fact that $\bar{P}_X X = 0$. Hence,

$$\begin{aligned} E(e'e) &= E(u'\bar{P}_X u) = E(\text{tr}\{u'\bar{P}_X u\}) = E(\text{tr}\{uu'\bar{P}_X\}) \\ &= \text{tr}(\sigma^2\bar{P}_X) = \sigma^2\text{tr}(\bar{P}_X) = \sigma^2(n-k) \end{aligned}$$

where the second equality follows from the fact that the trace of a scalar is a scalar. The third equality from the fact that $\text{tr}(ABC) = \text{tr}(CAB)$. The fourth equality from the fact that $E(trace) = trace\{E(.)\}$, and $E(uu') = \sigma^2 I_n$. The last equality from the fact that

$$\begin{aligned} \text{tr}(\bar{P}_X) &= \text{tr}(I_n) - \text{tr}(P_X) = n - \text{tr}(X(X'X)^{-1}X') \\ &= n - \text{tr}(X'X(X'X)^{-1}) = n - \text{tr}(I_k) = n - k. \end{aligned}$$

Hence, an unbiased estimator of $\text{var}(\widehat{\beta}_{OLS}) = \sigma^2(X'X)^{-1}$ is given by $s^2(X'X)^{-1}$.

So far we have shown that $\widehat{\beta}_{OLS}$ is BLUE. It can also be shown that it is consistent for β. In fact, taking probability limits of (7.5) as $n \to \infty$, one gets

$$\text{plim}(\widehat{\beta}_{OLS}) = \text{plim}(\beta) + \text{plim}(X'X/n)^{-1}(X'u/n) = \beta$$

The first equality uses the fact that the plim of a sum is the sum of the plims. The second equality follows from assumption 1 and the fact that plim of a product is the product of plims.

7.3 Partitioned Regression and the Frisch-Waugh-Lovell Theorem

In Chapter 4, we studied a useful property of least squares which allows us to interpret multiple regression coefficients as simple regression coefficients. This was called the residualing interpretation of multiple regression coefficients. In general, this property applies whenever the k regressors given by X can be separated into two sets of variables X_1 and X_2 of dimension $(n \times k_1)$ and $(n \times k_2)$ respectively, with $X = [X_1, X_2]$ and $k = k_1 + k_2$. The regression in equation (7.1) becomes a partitioned regression given by

$$y = X\beta + u = X_1\beta_1 + X_2\beta_2 + u \tag{7.8}$$

One may be interested in the least squares estimates of β_2 corresponding to X_2, but one has to control for the presence of X_1 which may include seasonal dummy variables or a time trend, see Frisch and Waugh (1933) and Lovell (1963)[1].

The OLS normal equations from (7.8) are as follows:

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} \widehat{\beta}_{1,OLS} \\ \widehat{\beta}_{2,OLS} \end{bmatrix} = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix} \tag{7.9}$$

These can be solved by partitioned inversion of the matrix on the left, see the Appendix to this chapter, or by solving two equations in two unknowns. Problem 2 asks the reader to verify that

$$\widehat{\beta}_{2,OLS} = (X_2'\bar{P}_{X_1}X_2)^{-1}X_2'\bar{P}_{X_1}y \tag{7.10}$$

where $\bar{P}_{X_1} = I_n - P_{X_1}$ and $P_{X_1} = X_1(X_1'X_1)^{-1}X_1'$. $\bar{P}_{X_1}$ is the orthogonal projection matrix of X_1 and $\bar{P}_{X_1}X_2$ generates the least squares residuals of each column of X_2 regressed on all the variables in X_1. In fact, if we denote by $\widetilde{X}_2 = \bar{P}_{X_1}X_2$ and $\widetilde{y} = \bar{P}_{X_1}y$, then (7.10) can be written as

$$\widehat{\beta}_{2,OLS} = (\widetilde{X}_2'\widetilde{X}_2)^{-1}\widetilde{X}_2'\widetilde{y} \tag{7.11}$$

using the fact that $\bar{P}_{X_1}$ is idempotent. This implies that $\widehat{\beta}_{2,OLS}$ can be obtained from the regression of $\widetilde{y}$ on $\widetilde{X}_2$. In words, the residuals from regressing y on X_1 are in turn regressed upon the residuals from each column of X_2 regressed on all the variables in X_1. This was illustrated in Chapter 4 with some examples. Following Davidson and MacKinnon (1993) we denote this result more formally as the Frisch-Waugh-Lovell (FWL) Theorem. In fact, if we premultiply (7.8) by $\bar{P}_{X_1}$ and use the fact that $\bar{P}_{X_1}X_1 = 0$, one gets

$$\bar{P}_{X_1}y = \bar{P}_{X_1}X_2\beta_2 + \bar{P}_{X_1}u \tag{7.12}$$

The FWL Theorem states that: (1) The least squares *estimates* of β_2 from equations (7.8) and (7.12) are numerically identical and (2) The least squares *residuals* from equations (7.8) and (7.12) are identical.

Using the fact that $\bar{\mathrm{P}}_{X_1}$ is idempotent, it immediately follows that, OLS on (7.12) yields $\widehat{\beta}_{2,OLS}$ as given by equation (7.10). Alternatively, one can start from equation (7.8) and use the result that

$$y = P_Xy + \bar{P}_Xy = X\widehat{\beta}_{OLS} + \bar{P}_Xy = X_1\widehat{\beta}_{1,OLS} + X_2\widehat{\beta}_{2,OLS} + \bar{P}_Xy \tag{7.13}$$

where $P_X = X(X'X)^{-1}X'$ and $\bar{P}_X = I_n - P_X$. Premultiplying equation (7.13) by $X_2'\bar{P}_{X_1}$ and using the fact that $\bar{P}_{X_1}X_1 = 0$, one gets

$$X_2'\bar{P}_{X_1}y = X_2'\bar{P}_{X_1}X_2\widehat{\beta}_{2,OLS} + X_2'\bar{P}_{X_1}\bar{P}_Xy \tag{7.14}$$

But, $P_{X_1}P_X = P_{X_1}$. Hence, $\bar{P}_{X_1}\bar{P}_X = \bar{P}_X$. Using this fact along with $\bar{P}_XX = \bar{P}_X[X_1, X_2] = 0$, the last term of equation (7.14) drops out yielding the result that $\widehat{\beta}_{2,OLS}$ from (7.14) is identical to the expression in (7.10). Note that no partitioned inversion was used in this proof. This proves part (1) of the FWL Theorem.

Also, premultiplying equation (7.13) by $\bar{P}_{X_1}$ and using the fact that $\bar{P}_{X_1}\bar{P}_X = \bar{P}_X$, one gets

$$\bar{P}_{X_1}y = \bar{P}_{X_1}X_2\widehat{\beta}_{2,OLS} + \bar{P}_X y \tag{7.15}$$

Now $\widehat{\beta}_{2,OLS}$ was shown to be numerically identical to the least squares estimate obtained from equation (7.12). Hence, the first term on the right hand side of equation (7.15) must be the fitted values from equation (7.12). Since the dependent variables are the same in equations (7.15) and (7.12), $\bar{P}_X y$ in equation (7.15) must be the least squares residuals from regression (7.12). But, $\bar{P}_X y$ is the least squares residuals from regression (7.8). Hence, the least squares residuals from regressions (7.8) and (7.12) are numerically identical. This proves part (2) of the FWL Theorem.

Several applications of the FWL Theorem will be given in this book. Problem 2 shows that if X_1 is the vector of ones indicating the presence of a constant in the regression, then regression (7.15) is equivalent to running $(y_i - \bar{y})$ on the set of variables in X_2 expressed as deviations from their respective sample means. Problem 3 shows that the FWL Theorem can be used to prove that including a dummy variable for one of the observations in the regression is equivalent to omitting that observation from the regression.

7.4 Maximum Likelihood Estimation

In Chapter 2, we introduced the method of maximum likelihood estimation which is based on specifying the distribution we are sampling from and writing the joint density of our sample. This joint density is then referred to as the likelihood function because it gives for a given set of parameters specifying the distribution, the probability of obtaining the observed sample. See Chapter 2 for several examples. For the regression equation, specifying the distribution of the disturbances in turn specifies the likelihood function. These disturbances could be Poisson, Exponential, Normal, etc. Once this distribution is chosen, the likelihood function is maximized and the MLE of the regression parameters are obtained. Maximum likelihood estimators are desirable because they are (1) consistent under fairly general conditions,[2] (2) asymptotically normal, (3) asymptotically efficient and (4) invariant to reparameterizations of the model[3]. Some of the undesirable properties of MLE are that (1) it requires explicit distributional assumptions on the disturbances, and (2) their finite sample properties can be quite different from their asymptotic properties. For example, MLE can be biased even though they are consistent, and their covariance estimates can be misleading for small samples. In this section, we derive the MLE under normality of the disturbances.

The Normality Assumption: $u \sim N(0, \sigma^2 I_n)$**.** This additional assumption allows us to derive distributions of estimators and other random variables. This is important for constructing confidence intervals and tests of hypotheses. In fact using (7.5) one can easily see that $\widehat{\beta}_{OLS}$ is a linear combination of the u's. But, a linear combination of normal random variables is itself a normal random variable. Hence, $\widehat{\beta}_{OLS}$ is $N(\beta, \sigma^2(X'X)^{-1})$. Similarly y is $N(X\beta, \sigma^2 I_n)$ and e is $N(0, \sigma^2\bar{P}_X)$. Moreover, we can write the joint probability density function of the u's as $f(u_1, u_2, \ldots, u_n; \beta, \sigma^2) = (1/2\pi\sigma^2)^{n/2}\exp(-u'u/2\sigma^2)$. To get the likelihood function we make the transformation $u = y - X\beta$ and note that the Jacobian of the transformation is one. Hence

$$f(y_1, y_2, \ldots, y_n; \beta, \sigma^2) = (1/2\pi\sigma^2)^{n/2}\exp\{-(y - X\beta)'(y - X\beta)/2\sigma^2\} \tag{7.16}$$

Taking the log of this likelihood, we get

$$\log L(\beta, \sigma^2) = -(n/2)\log(2\pi\sigma^2) - (y - X\beta)'(y - X\beta)/2\sigma^2 \tag{7.17}$$

Maximizing this likelihood with respect to β and σ^2 one gets the maximum likelihood estimators (MLE). Let $\theta = \sigma^2$ and $Q = (y - X\beta)'(y - X\beta)$, then

$$\begin{aligned} \frac{\partial log L(\beta,\theta)}{\partial \beta} &= \frac{2X'y - 2X'X\beta}{2\theta} \\ \frac{\partial log L(\beta,\theta)}{\partial \theta} &= \frac{Q}{2\theta^2} - \frac{n}{2\theta} \end{aligned}$$

Setting these first-order conditions equal to zero, one gets

$$\widehat{\beta}_{MLE} = \widehat{\beta}_{OLS} \quad \text{and} \quad \widehat{\theta} = \widehat{\sigma}^2_{MLE} = Q/n = RSS/n = e'e/n.$$

Intuitively, only the second term in the log likelihood contains β and that term (without the negative sign) has already been minimized with respect to β in (7.2) giving us the OLS estimator. Note that $\widehat{\sigma}^2_{MLE}$ differs from s^2 only in the degrees of freedom. It is clear that $\widehat{\beta}_{MLE}$ is unbiased for β while $\widehat{\sigma}^2_{MLE}$ is not unbiased for σ^2. Substituting these MLE's into (7.17) one gets the maximum value of $\log L$ which is

$$\begin{aligned} \log L(\widehat{\beta}_{MLE}, \widehat{\sigma}^2_{MLE}) &= -(n/2)\log(2\pi\widehat{\sigma}^2_{MLE}) - e'e/2\widehat{\sigma}^2_{MLE} \\ &= -(n/2)\log(2\pi) - (n/2)\log(e'e/n) - n/2 \\ &= \text{constant} - (n/2)\log(e'e). \end{aligned}$$

In order to get the Cramér-Rao lower bound for the unbiased estimators of β and σ^2 one first computes the information matrix

$$I(\beta, \sigma^2) = -E \begin{bmatrix} \partial^2 \log L/\partial\beta\partial\beta' & \partial^2 \log L/\partial\beta\partial\sigma^2 \\ \partial^2 \log L/\partial\sigma^2\partial\beta' & \partial^2 \log L/\partial\sigma^2\partial\sigma^2 \end{bmatrix} \tag{7.18}$$

Recall, that $\theta = \sigma^2$ and $Q = (y - X\beta)'(y - X\beta)$. It is easy to show (see problem 4) that

$$\frac{\partial^2 \log L(\beta,\theta)}{\partial\beta\partial\theta} = \frac{1}{2\theta^2}\frac{\partial Q}{\partial\beta} \quad \text{and} \quad \frac{\partial^2 \log L(\beta,\theta)}{\partial\theta\partial\beta} = \frac{-X'(y - X\beta)}{\theta^2}$$

Therefore,

$$E\left(\frac{\partial^2 \log L(\beta,\theta)}{\partial\theta\partial\beta}\right) = \frac{-E(X'u)}{\theta^2} = 0$$

Also

$$\frac{\partial^2 \log L(\beta,\theta)}{\partial\beta\partial\beta'} = \frac{-X'X}{\theta} \quad \text{and} \quad \frac{\partial^2 \log L(\beta,\theta)}{\partial\theta^2} = \frac{-4Q}{4\theta^3} + \frac{2n}{4\theta^2} = \frac{-Q}{\theta^3} + \frac{n}{2\theta^2}$$

so that

$$E\left(\frac{\partial^2 \log L(\beta,\theta)}{\partial\theta^2}\right) = \frac{-n\theta}{\theta^3} + \frac{n}{2\theta^2} = \frac{-2n + n}{2\theta^2} = \frac{-n}{2\theta^2}$$

using the fact that $E(Q) = n\sigma^2 = n\theta$. Hence,

$$I(\beta, \sigma^2) = \begin{bmatrix} X'X/\sigma^2 & 0 \\ 0 & n/2\sigma^4 \end{bmatrix} \tag{7.19}$$

The information matrix is block-diagonal between β and σ^2. This is an important property for regression models with normal disturbances. It implies that the Cramér-Rao lower bound is

$$I^{-1}(\beta, \sigma^2) = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & 2\sigma^4/n \end{bmatrix} \tag{7.20}$$

Note that $\widehat{\beta}_{MLE} = \widehat{\beta}_{OLS}$ attains the Cramér-Rao lower bound. Under normality, $\widehat{\beta}_{OLS}$ is MVU (minimum variance unbiased). This is best among all unbiased estimators not only *linear* unbiased estimators. By assuming more (in this case normality) we get more (MVU rather than BLUE)[4].

Problem 5 derives the variance of s^2 under normality of the disturbances. This is found to be $2\sigma^4/(n-k)$. This means that s^2 does not attain the Cramér-Rao lower bound. However, following the theory of complete sufficient statistics one can show that both $\widehat{\beta}_{OLS}$ and s^2 are MVU for their respective parameters and therefore both are small sample efficient. Note also that $\widehat{\sigma}^2_{MLE}$ is biased, therefore it is not meaningful to compare its variance to the Cramér-Rao lower bound. There is a trade-off between bias and variance in estimating σ^2. Problem 6 looks at all estimators of σ^2 of the type $e'e/r$ and derives r such that the mean squared error (MSE) is minimized. The choice of r turns out to be $(n-k+2)$.

We found the distribution of $\widehat{\beta}_{OLS}$, now we derive the distribution of s^2. In order to do that we need a result from matrix algebra, which is stated without proof, see Graybill (1961).

Lemma 1: For every symmetric idempotent matrix A of rank r, there exists an orthogonal matrix P such that $P'AP = J_r$ where J_r is a diagonal matrix with the first r elements equal to one and the rest equal to zero.

We use this lemma to show that the RSS/σ^2 is a chi-squared with $(n-k)$ degrees of freedom. To see this note that $e'e/\sigma^2 = u'\bar{P}_X u/\sigma^2$ and that $\bar{P}_X$ is symmetric and idempotent of rank $(n-k)$. Using the lemma there exists a matrix P such that $P'\bar{P}_X P = J_{n-k}$ is a diagonal matrix with the first $(n-k)$ elements on the diagonal equal to 1 and the last k elements equal to zero. Now make the change of variable $v = P'u$. This makes $v \sim N(0, \sigma^2 I_n)$ since the v's are linear combinations of the u's and $P'P = I_n$. Replacing u by v in RSS/σ^2 we get

$$v'P\bar{P}_X Pv/\sigma^2 = v'J_{n-k}v/\sigma^2 = \textstyle\sum_{i=1}^{n-k} v_i^2/\sigma^2$$

where the last sum is only over $i = 1, 2, \ldots, n-k$. But, the v's are independent identically distributed $N(0, \sigma^2)$, hence v_i^2/σ^2 is the square of a standardized $N(0,1)$ random variable which is distributed as a χ^2_1. Moreover, the sum of independent χ^2 random variables is a χ^2 random variable with degrees of freedom equal to the sum of the respective degrees of freedom. Hence, RSS/σ^2 is distributed as χ^2_{n-k}.

The beauty of the above result is that it applies to all quadratic forms $u'Au$ where A is symmetric and idempotent. We will use this result again in the test of hypotheses section.

7.5 Prediction

Let us now predict T_o periods ahead. Those new observations are assumed to satisfy (7.1). In other words

$$y_o = X_o\beta + u_o \tag{7.21}$$

What is the Best Linear Unbiased Predictor (BLUP) of $E(y_o)$? From (7.21), $E(y_o) = X_o\beta$ which is a linear combination of the β's. Using the Gauss-Markov result $\widehat{y}_o = X_o\widehat{\beta}_{OLS}$ is BLUE for $X_o\beta$ and the variance of this predictor of $E(y_o)$ is $X_o\text{var}(\widehat{\beta}_{OLS})X_o' = \sigma^2 X_o(X'X)^{-1}X_o'$. But, what if we are interested in the predictor for y_o? The best predictor of u_o is zero, so the predictor for y_o is still $\widehat{y}_o$ but its MSE is

$$\begin{aligned} &E(\widehat{y}_o - y_o)(\widehat{y}_o - y_o)' = E\{X_o(\widehat{\beta}_{OLS} - \beta) - u_o\}\{X_o(\widehat{\beta}_{OLS} - \beta) - u_o\}' \\ &= X_o\text{var}(\widehat{\beta}_{OLS})X_o' + \sigma^2 I_{T_o} - 2\text{cov}\{X_o(\widehat{\beta}_{OLS} - \beta), u_o\} \\ &= \sigma^2 X_o(X'X)^{-1}X_o' + \sigma^2 I_{T_o} \end{aligned} \tag{7.22}$$

the last equality follows from the fact that $(\widehat{\beta}_{OLS} - \beta) = (X'X)^{-1}X'u$ and u_o have zero covariance. The latter holds because u_o and u have zero covariance. Intuitively this says that the future T_o disturbances are not correlated with the current sample disturbances.

Therefore, the predictor of the average consumption of a $\$20,000$ income household is the same as the predictor of consumption of a specific household whose income is $\$20,000$. The difference is not in the predictor itself but in the MSE attached to it. The latter MSE being larger.

Salkever (1976) suggested a simple way to compute these forecasts and their standard errors. The basic idea is to augment the usual regression in (7.1) with a matrix of observation-specific dummies, i.e., a dummy variable for each period where we want to forecast:

$$\begin{bmatrix} y \\ y_o \end{bmatrix} = \begin{bmatrix} X & 0 \\ X_o & I_{T_o} \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} + \begin{bmatrix} u \\ u_o \end{bmatrix} \tag{7.23}$$

or

$$y^* = X^*\delta + u^* \tag{7.24}$$

where $\delta' = (\beta', \gamma')$. X^* has in its second part a matrix of dummy variables one for each of the T_o periods for which we are forecasting. Since these T_o observations do not serve in the estimation, problem 7 asks the reader to verify that OLS on (7.23) yields $\widehat{\delta}' = (\widehat{\beta}', \widehat{\gamma}')$ where $\widehat{\beta} = (X'X)^{-1}X'y$, $\widehat{\gamma} = y_o - \widehat{y}_o$, and $\widehat{y}_o = X_o\widehat{\beta}$. In other words, OLS on (7.23) yields the OLS estimate of β without the T_o observations, and the coefficients of the T_o dummies are the forecast errors. This also means that the first n residuals are the usual OLS residuals $e = y - X\widehat{\beta}$ based on the first n observations, whereas the next T_o residuals are all zero. Therefore, $s^{*2} = s^2 = e'e/(n-k)$, and the variance covariance matrix of $\widehat{\delta}$ is given by

$$s^2(X^{*\prime}X^*)^{-1} = s^2 \begin{bmatrix} (X'X)^{-1} & \\ & [I_{T_o} + X_o(X'X)^{-1}X_o'] \end{bmatrix} \tag{7.25}$$

and the off-diagonal elements are of no interest. This means that the regression package gives the estimated variance of $\widehat{\beta}$ and the estimated variance of the forecast error in one stroke. Note that if the forecasts rather than the forecast errors are needed, one can replace y_o by zero, and I_{T_o} by $-I_{T_o}$ in (7.23). The resulting estimate of γ will be $\widehat{y}_o = X_o\widehat{\beta}$, as required. The variance of this forecast will be the same as that given in (7.25), see problem 7.

7.6 Confidence Intervals and Test of Hypotheses

We start by constructing a confidence interval for any linear combination of β, say $c'\beta$. We know that $c'\widehat{\beta}_{OLS} \sim N(c'\beta, \sigma^2 c'(X'X)^{-1}c)$ and it is a scalar. Hence,

$$z_{obs} = (c'\widehat{\beta}_{OLS} - c'\beta)/\sigma(c'(X'X)^{-1}c)^{1/2} \tag{7.26}$$

is a standardized $N(0,1)$ random variable. Replacing σ by s is equivalent to dividing z_{obs} by the square root of a χ^2 random variable divided by its degrees of freedom. The latter random variable is $(n-k)s^2/\sigma^2 = RSS/\sigma^2$ which was shown to be a χ^2_{n-k}. Problem 8 shows that z_{obs} and RSS/σ^2 are independent. This means that

$$t_{obs} = (c'\widehat{\beta}_{OLS} - c'\beta)/s(c'(X'X)^{-1}c)^{1/2} \tag{7.27}$$

is a $N(0,1)$ random variable divided by the square root of an independent $\chi^2_{n-k}/(n-k)$. This is a t-statistic with $(n-k)$ degrees of freedom. Hence, a $100(1-\alpha)\%$ confidence interval for $c'\beta$ is

$$c'\widehat{\beta}_{OLS} \pm t_{\alpha/2}s(c'(X'X)^{-1}c)^{1/2} \tag{7.28}$$

Example: Let us say we are predicting one year ahead so that $T_o = 1$ and x_o is a $(1 \times k)$ vector of next year's observations on the exogenous variables. The $100(1-\alpha)$ confidence interval for next year's forecast of y_o will be $\widehat{y}_o \pm t_{\alpha/2}s(1 + x_o'(X'X)^{-1}x_o)^{1/2}$. Similarly (7.28) allows us to construct confidence intervals or test any single hypothesis on any single β_j (again by picking c to have 1 in its j-th position and zero elsewhere). In this case we get the usual t-statistic reported in any regression package. More importantly, this allows us to test any hypothesis concerning any linear combination of the β's, e.g., testing that the sum of coefficients of input variables in a Cobb-Douglas production function is equal to one. This is known as a test for constant returns to scale, see Chapter 4.

7.7 Joint Confidence Intervals and Test of Hypotheses

We have learned how to test any single hypothesis involving any linear combination of the β's. But what if we are interested in testing two or three or more hypotheses involving linear combinations of the β's. For example, testing that $\beta_2 = \beta_4 = 0$, i.e., that variables X_2 and X_4 are not significant in the model. This can be written as $c_2'\beta = c_4'\beta = 0$ where c_j' is a row vector of zeros with a one in the j-th position. In order to test these two hypotheses simultaneously, we rearrange these restrictions on the β's in matrix form $R\beta = 0$ where $R' = [c_2, c_4]$. In a similar fashion, we can rearrange g restrictions on the β's into this matrix R which will now be of dimension $(g \times k)$. Also these restrictions need not be of the form $R\beta = 0$ and can be of the more general form $R\beta = r$ where r is a $(g \times 1)$ vector of constants. For example, $\beta_1 + \beta_2 = 1$ and $3\beta_3 + 2\beta_4 = 5$ are two such restrictions. Since $R\beta$ is a collection of linear combinations of the β's, the BLUE of these is $R\widehat{\beta}_{OLS}$ and the latter is distributed $N(R\beta, \sigma^2 R(X'X)^{-1}R')$. Standardization of the form encountered with the scalar $c'\beta$ gives us the following:

$$(R\widehat{\beta}_{OLS} - R\beta)'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - R\beta)/\sigma^2 \tag{7.29}$$

rather than divide by the variance we multiply by its inverse, and since we divided by the variance rather than the standard deviation we square the numerator which means in vector form premultiplying by its transpose. Problem 9 replaces the matrix R by the vector c' and shows that (7.29) reduces to the square of the z-statistic observed in (7.26). This also proves that the resulting statistic is distributed as χ^2_1. But, what is the distribution of (7.29)? The trick is to write it in terms of the original disturbances, i.e.,

$$u'X(X'X)^{-1}R'[R(X'X)^{-1}R]^{-1}R(X'X)^{-1}X'u/\sigma^2 \tag{7.30}$$

where $(R\widehat{\beta}_{OLS} - R\beta)$ is replaced by $R(X'X)^{-1}X'u$. Note that (7.30) is quadratic in the disturbances u of the form $u'Au/\sigma^2$. Problem 10 shows that A is symmetric and idempotent and of rank g. Applying the same proof as given below lemma 1 we get the result that (7.30) is distributed as χ^2_g. Again σ^2 is unobserved, so we divide by $(n-k)s^2/\sigma^2$ which is χ^2_{n-k}. This becomes a ratio of two χ^2's random variables. If we divide the numerator and denominator χ^2's by their respective degrees of freedom and prove that they are independent (see problem 11) the resulting statistic

$$(R\widehat{\beta}_{OLS} - r)'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - r)/gs^2 \tag{7.31}$$

is distributed under the null $R\beta = r$ as an $F(g, n-k)$.

7.8 Restricted MLE and Restricted Least Squares

Maximizing the likelihood function given in (7.16) subject to $R\beta = r$ is equivalent to minimizing the residual sum of squares subject to $R\beta = r$. Forming the Lagrangian function

$$\Psi(\beta, \mu) = (y - X\beta)'(y - X\beta) + 2\mu'(R\beta - r) \tag{7.32}$$

and differentiating with respect to β and μ one gets

$$\partial\Psi(\beta, \mu)/\partial\beta = -2X'y + 2X'X\beta + 2R'\mu = 0 \tag{7.33}$$

$$\partial\Psi(\beta, \mu)/\partial\mu = 2(R\beta - r) = 0 \tag{7.34}$$

Solving for μ, we premultiply (7.33) by $R(X'X)^{-1}$ and use (7.34)

$$\widehat{\mu} = [R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - r) \tag{7.35}$$

Substituting (7.35) in (7.33) we get

$$\widehat{\beta}_{RLS} = \widehat{\beta}_{OLS} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - r) \tag{7.36}$$

The restricted least squares estimator of β differs from that of the unrestricted OLS estimator by the second term in (7.36) with the term in parentheses showing the extent to which the unrestricted OLS estimator satisfies the constraint. Problem 12 shows that $\widehat{\beta}_{RLS}$ is biased unless the restriction $R\beta = r$ is satisfied. However, its variance is always less than that of $\widehat{\beta}_{OLS}$. This brings in the trade-off between bias and variance and the MSE criteria which was discussed in Chapter 2.

The Lagrange Multiplier estimator $\widehat{\mu}$ is distributed $N(0, \sigma^2[R(X'X)^{-1}R']^{-1})$ under the null hypothesis. Therefore, to test $\mu = 0$, we use

$$\widehat{\mu}'[R(X'X)^{-1}R']\widehat{\mu}/\sigma^2 = (R\widehat{\beta}_{OLS} - r)'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - r)/\sigma^2 \tag{7.37}$$

Since μ measures the cost of imposing the restriction $R\beta = r$, it is no surprise that the right hand side of (7.37) was already encountered in (7.29) and is distributed as χ^2_g.

7.9 Likelihood Ratio, Wald and Lagrange Multiplier Tests

Before we go into the derivations of these three classical tests for the null hypothesis $H_0; R\beta = r$, it is important for the reader to review the intuitive graphical explanation of these tests given in Chapter 2.

The Likelihood Ratio test of $H_0; R\beta = r$ is based upon the ratio $\lambda = \max\ell_r/\max\ell_u$, where $\max\ell_u$ and $\max\ell_r$ are the maximum values of the unrestricted and restricted likelihoods, respectively. Let us assume for simplicity that σ^2 is *known*, then

$$\max\ell_u = (1/2\pi\sigma^2)^{n/2}\exp\{-(y - X\widehat{\beta}_{MLE})'(y - X\widehat{\beta}_{MLE})/2\sigma^2\}$$

where $\widehat{\beta}_{MLE} = \widehat{\beta}_{OLS}$. Denoting the unrestricted residual sum of squares by URSS, we have

$$\max\ell_u = (1/2\pi\sigma^2)^{n/2}\exp\{-URSS/2\sigma^2\}$$

Similarly, $\max\ell_r$ is given by

$$\max\ell_r = (1/2\pi\sigma^2)^{n/2}\exp\{-(y - X\widehat{\beta}_{RMLE})'(y - X\widehat{\beta}_{RMLE})/2\sigma^2\}$$

where $\widehat{\beta}_{RMLE} = \widehat{\beta}_{RLS}$. Denoting the restricted residual sum of squares by RRSS, we have

$$\max\ell_r = (1/2\pi\sigma^2)^{n/2}\exp\{-RRSS/2\sigma^2\}$$

Therefore, $-2\log\lambda = (RRSS - URSS)/\sigma^2$. Let us find the relationship between these residual sums of squares.

$$\begin{aligned} e_r &= y - X\widehat{\beta}_{RLS} = y - X\widehat{\beta}_{OLS} - X(\widehat{\beta}_{RLS} - \widehat{\beta}_{OLS}) = e - X(\widehat{\beta}_{RLS} - \widehat{\beta}_{OLS}) \qquad (7.38)\\ e_r'e_r &= e'e + (\widehat{\beta}_{RLS} - \widehat{\beta}_{OLS})'X'X(\widehat{\beta}_{RLS} - \widehat{\beta}_{OLS}) \end{aligned}$$

where e_r denotes the restricted residuals and $e_r'e_r$ the RRSS. The cross-product terms drop out because $X'e = 0$. Substituting the value of $(\widehat{\beta}_{RLS} - \widehat{\beta}_{OLS})$ from (7.36) into (7.38), we get:

$$RRSS - URSS = (R\widehat{\beta}_{OLS} - r)'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - r) \qquad (7.39)$$

It is now clear that $-2\log\lambda$ is the right hand side of (7.39) divided by σ^2. In fact, this Likelihood Ratio (LR) statistic is the same as that given in (7.37) and (7.29). Under the null hypothesis $R\beta = r$, this was shown to be a χ^2_g.

The Wald test of $R\beta = r$ is based upon the unrestricted estimator and the extent of which it satisfies the restriction. More formally, if $r(\beta) = 0$ denote the vector of g restrictions on β and $R(\widehat{\beta}_{MLE})$ denotes the $(g \times k)$ matrix of partial derivatives $\partial r(\beta)/\partial\beta'$ evaluated at $\widehat{\beta}_{MLE}$, then the Wald statistic is given by

$$W = r(\widehat{\beta}_{MLE})'[R(\widehat{\beta}_{MLE})I(\widehat{\beta}_{MLE})^{-1}R(\widehat{\beta}_{MLE})']^{-1}r(\widehat{\beta}_{MLE}) \qquad (7.40)$$

where $I(\beta) = -E(\partial^2\log L/\partial\beta\partial\beta')$. In this case, $r(\beta) = R\beta - r, R(\widehat{\beta}_{MLE}) = R$ and $I(\widehat{\beta}_{MLE}) = (X'X)/\sigma^2$ as seen in (7.19). Therefore,

$$W = (R\widehat{\beta}_{MLE} - r)'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{MLE} - r)/\sigma^2 \qquad (7.41)$$

which is the same as the LR statistic[5].

The Lagrange Multiplier test is based upon the restricted estimator. In section 7.8, we derived the restricted estimator and the estimated Lagrange Multiplier $\widehat{\mu}$. The Lagrange Multiplier μ is the cost or shadow price of imposing the restrictions $R\beta = r$. If these restrictions are true, one would expect the estimated Lagrange Multiplier $\widehat{\mu}$ to have mean zero. Therefore, a test for the null hypothesis that $\mu = 0$, is called the LM test and the corresponding test statistic is given in equation (7.37). Alternatively, one can derive the LM test as a score' test based on the score or the first derivative of the log-likelihood function i.e., $S(\beta) = \partial \log L/\partial\beta$. The score is zero for the unrestricted MLE, and the score test is based upon the departure of $S(\beta)$, evaluated at the restricted estimator $\widehat{\beta}_{RMLE}$, from zero. In this case, the score form of the LM statistic is given by

$$LM = S(\widehat{\beta}_{RMLE})' I(\widehat{\beta}_{RMLE})^{-1} S(\widehat{\beta}_{RMLE}) \tag{7.42}$$

For our model, $S(\beta) = (X'y - X'X\beta)/\sigma^2$ and from equation (7.36) we have

$$\begin{aligned} S(\widehat{\beta}_{RMLE}) &= X'(y - X\widehat{\beta}_{RMLE})/\sigma^2 \\ &= \{X'y - X'X\widehat{\beta}_{OLS} + R'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - r)\}/\sigma^2 \\ &= R'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - r)/\sigma^2 \end{aligned}$$

Using (7.20), one gets $I^{-1}(\widehat{\beta}_{RMLE}) = \sigma^2(X'X)^{-1}$. Therefore, the score form of the LM test becomes

$$\begin{aligned} LM &= (R\widehat{\beta}_{OLS} - r)'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - r)/\sigma^2 \\ &= (R\widehat{\beta}_{OLS} - r)'[R(X'X)^{-1}R']^{-1}(R\widehat{\beta}_{OLS} - r)/\sigma^2 \end{aligned} \tag{7.43}$$

This is numerically identical to the LM test derived in equation (7.37) and to the W and LR statistics derived above. Note that $\mathrm{S}(\widehat{\beta}_{RMLE}) = R'\widehat{\mu}/\sigma^2$ from (7.35), so it is clear why the Score and the Lagrangian Multiplier tests are identical.

The score form of the LM test can also be obtained as a by-product of an artificial regression. In fact, $S(\beta)$ evaluated as $\widehat{\beta}_{RMLE}$ is given by

$$S(\widehat{\beta}_{RMLE}) = X'(y - X\widehat{\beta}_{RMLE})/\sigma^2$$

where $y - X\widehat{\beta}_{RMLE}$ is the vector of restricted residuals. If H_0 is true, then this converges asymptotically to u and the asymptotic variance of the vector of scores becomes $(X'X)/\sigma^2$. The score test is then based upon

$$(y - X\widehat{\beta}_{RMLE})'X(X'X)^{-1}X'(y - X\widehat{\beta}_{RMLE})/\sigma^2 \tag{7.44}$$

This expression is the explained sum of squares from the artificial regression of $(y - X\widehat{\beta}_{RMLE})/\sigma$ on X. To see that this is exactly identical to the LM test in equation (7.37), recall from equation (7.33) that $R'\widehat{\mu} = X'(y - X\widehat{\beta}_{RMLE})$ and substituting this expression for $R'\widehat{\mu}$ on the left hand side of equation (7.37) we get equation (7.44). In practice, σ^2 is estimated by $\widetilde{s}^2$ the Mean Square Error of the restricted regression. This is an example of the Gauss-Newton Regression which will be discussed in Chapter 8.

An alternative approach to testing H_0, is to estimate the restricted and unrestricted models and compute the following F-statistic

$$F_{obs} = \frac{(RRSS - URSS)/g}{URSS/(n-k)} \tag{7.45}$$

This statistic is known in the econometric literature as the Chow (1960) test and was encountered in Chapter 4. Note that from equation (7.39), if we divide the numerator by σ^2 we get a χ^2_g statistic divided by its degrees of freedom. Also, using the fact that $(n-k)s^2/\sigma^2$ is χ^2_{n-k}, the denominator divided by σ^2 is a χ^2_{n-k} statistic divided by its degrees of freedom. Problem 11 shows independence of the numerator and denominator and completes the proof that F_{obs} is distributed $F(g, n-k)$ under H_0.

Chow's (1960) Test for Regression Stability

Chow (1960) considered the problem of testing the equality of two sets of regression coefficients

$$y_1 = X_1\beta_1 + u_1 \quad \text{and} \quad y_2 = X_2\beta_2 + u_2 \tag{7.46}$$

where X_1 is $n_1 \times k$ and X_2 is $n_2 \times k$ with n_1 and $n_2 > k$. In this case, the unrestricted regression can be written as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{7.47}$$

under the null hypothesis $H_0; \beta_1 = \beta_2 = \beta$, the restricted model becomes

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{7.48}$$

The URSS and the RRSS are obtained from these two regressions by stacking the $n_1 + n_2$ observations. It is easy to show that the $URSS = e_1'e_1 + e_2'e_2$ where e_1 is the OLS residuals from y_1 on X_1 and e_2 is the OLS residuals from y_2 on X_2. In other words, the URSS is the sum of two residual sums of squares from the separate regressions, see problem 13. The Chow F-statistic given in equation (7.45) has k and $(n_1 + n_2 - 2k)$ degrees of freedom, respectively. Equivalently, one can obtain this Chow F-statistic from running

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta_1 + \begin{bmatrix} 0 \\ X_2 \end{bmatrix} (\beta_2 - \beta_1) + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{7.49}$$

Note that the second set of explanatory variables whose coefficients are $(\beta_2 - \beta_1)$ are interaction variables obtained by multiplying each independent variable in equation (7.48) by a dummy variable, say D_2, that takes on the value 1 if the observation is from the second regression and 0 if it is from the first regression. A test for H_0; $\beta_1 = \beta_2$ becomes a joint test of significance for the coefficients of these interaction variables. Gujarati (1970) points out that this dummy variable approach has the additional advantage of giving the estimates of $(\beta_2 - \beta_1)$ and their t-statistics. If the Chow F-test rejects stability, these individual interaction dummy variable coefficients may point to the source of instability. Of course, one has to be careful with the interpretation of these individual t-statistics, after all they can all be insignificant with the joint F-statistic still being significant, see Maddala (1992).

In case one of the two regressions does not have sufficient observations to estimate a separate regression say $n_2 < k$, then one can proceed by running the regression on the full data set to get the RRSS. This is the restricted model because the extra n_2 observations are assumed to be generated by the same regression as the first n_1 observations. The URSS is the residual sums

of squares based only on the longer period (n_1 observations). In this case, the Chow F-statistic given in equation (7.45) has n_2 and $n_1 - k$ degrees of freedom, respectively. This is known as Chow's *predictive test* since it tests whether the shorter n_2 observations are different from their predictions using the model with the longer n_1 observations. This predictive test can be performed with dummy variables as follows: Introduce n_2 observation specific dummies, one for each of the observations in the second regression. Test the joint significance of these n_2 dummy variables. Salkever's (1976) result applies and each dummy variable will have as its estimated coefficient the prediction error with its corresponding standard error and its t-statistic. Once, again, the individual dummies may point out possible outliers, but it is their joint significance that is under question.

The W, LR and LM Inequality

We have shown that the $LR = W = LM$ for linear restrictions if the log-likelihood is quadratic. However, this is not necessarily the case for more general situations. In fact, in the next chapter where we consider more general variance covariance structure on the disturbances, estimating this variance-covariance matrix destroys this equality and may lead to conflict in hypotheses testing as noted by Berndt and Savin (1977). In this case, $W \geq LR \geq LM$. See also the problems at the end of this chapter. The LR, W and LM tests are based on the *efficient* MLE. When *consistent* rather than *efficient* estimators are used, an alternative way of constructing the score-type test is known as Neyman's $C(\alpha)$. For details, see Bera and Permaratne (2001).

Although, these three tests are asymptotically equivalent, one test may be more convenient than another for a particular problem. For example, when the model is linear but the restriction is nonlinear, the unrestricted model is easier to estimate than the restricted model. So the Wald test suggests itself in that it relies only on the unrestricted estimator. Unfortunately, the Wald test has a drawback that the LR and LM test do not have. In finite samples, the Wald test is not invariant to testing two algebraically equivalent formulations of the nonlinear restriction. This fact has been pointed out in the econometric literature by Gregory and Veall (1985, 1986) and Lafontaine and White (1986). In what follows, we review some of Gregory and Veall's (1985) findings:

Consider the linear regression with two regressors

$$y_t = \beta_o + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t \tag{7.50}$$

where the u_t's are IIN$(0, \sigma^2)$, and the nonlinear restriction $\beta_1\beta_2 = 1$. Two algebraically equivalent formulation of the null hypothesis are: H^A; $r^A(\beta) = \beta_1 - 1/\beta_2 = 0$, and H^B; $r^B(\beta) = \beta_1\beta_2 - 1 = 0$. The unrestricted maximum likelihood estimator is $\widehat{\beta}_{OLS}$ and the Wald statistic given in (7.40) is

$$W = r(\widehat{\beta}_{OLS})'[R(\widehat{\beta}_{OLS})\widehat{V}(\widehat{\beta}_{OLS})R'(\widehat{\beta}_{OLS})]^{-1}r(\widehat{\beta}_{OLS}) \tag{7.51}$$

where $\widehat{V}(\widehat{\beta}_{OLS})$ is the usual estimated variance-covariance matrix of $\widehat{\beta}_{OLS}$. Problem 19 asks the reader to verify that the Wald statistics corresponding to H_A and H_B using (7.51) are

$$W^A = (\widehat{\beta}_1\widehat{\beta}_2 - 1)^2/(\widehat{\beta}_2^2 v_{11} + 2v_{12} + v_{22}/\widehat{\beta}_2^2) \tag{7.52}$$

and

$$W^B = (\widehat{\beta}_1\widehat{\beta}_2 - 1)^2/(\widehat{\beta}_2^2 v_{11} + 2\widehat{\beta}_1\widehat{\beta}_2 v_{12} + \widehat{\beta}_1^2 v_{22}) \tag{7.53}$$

where the v_{ij}'s are the elements of $\widehat{V}(\widehat{\beta}_{OLS})$ for $i, j = 0, 1, 2$. These Wald statistics are clearly not identical, and other algebraically equivalent formulations of the null hypothesis can be generated with correspondingly different Wald statistics. Monte Carlo experiments were performed with 1000 replications on the model given in (7.50) with various values for β_1 and β_2, and for a sample size $n = 20, 30, 50, 100, 500$. The experiments were run when the null hypothesis is true and when it is false. For $n = 20$ and $\beta_1 = 10$, $\beta_2 = 0.1$, so that H_0 is satisfied, W^A rejects the null when it is true 293 times out of a 1000, while W^B rejects the null 65 times out of a 1000. At the 5% level one would expect 50 rejections with a 95% confidence interval $[36, 64]$. Both W^A and W^B reject too often but W^A performs worse than W^B. When n is increased to 500, W^A rejects 78 times while W^B rejects 39 times out of a 1000. W^A still rejects too often although its performance is better than that for $n = 20$, while W^B performs well and is within the 95% confidence region. When $n = 20$, $\beta_1 = 1$ and $\beta_2 = 0.5$, so that H_0 is *not* satisfied, W^A rejects the null when it is false 65 times out of a 1000 whereas W^B rejects it 584 times out of a 1000. For $n = 500$, both test statistics reject the null in 1000 out of 1000 times. Even in cases where the empirical sizes of the tests appear similar, see Table 1 of Gregory and Veall (1985), in particular the case where $\beta_1 = \beta_2 = 1$, Gregory and Veall find that W^A and W^B are in conflict about 5% of the time for $n = 20$, and this conflict drops to 0.5% at $n = 500$. Problem 20 asks the reader to derive four Wald statistics corresponding to four algebraically equivalent formulations of the *common factor* restriction analyzed by Hendry and Mizon (1978). Gregory and Veall (1986) give Monte Carlo results on the performance of these Wald statistics for various sample sizes. Once again they find conflict among these tests even when their empirical sizes appear to be similar. Also, the differences among the Wald statistics are much more substantial, and persist even when n is as large as 500.

Lafontaine and White (1985) consider a simple regression

$$y = \alpha + \beta x + \gamma z + u$$

where y is log of per capita consumption of textiles, x is log of per capita real income and z is log of relative prices of textiles, with the data taken from Theil (1971, p.102). The estimated equation is:

$$\begin{array}{rccc} \widehat{y} = & 1.37 & + 1.14x & - 0.83z \\ & (0.31) & (0.16) & (0.04) \end{array}$$

with $\widehat{\sigma}^2 = 0.0001833$, and $n = 17$, with standard errors shown in parentheses. Consider the null hypothesis H_0; $\beta = 1$. Algebraically equivalent formulations of H_0 are H_k; $\beta^k = 1$ for any exponent k. Applying (7.40) with $r(\beta) = \beta^k - 1$ and $R(\beta) = k\beta^{k-1}$, one gets the Wald statistic

$$W_k = (\widehat{\beta}^k - 1)^2 / [(k\widehat{\beta}^{k-1})^2 V(\widehat{\beta})] \tag{7.54}$$

where $\widehat{\beta}$ is the OLS estimate of β and $V(\widehat{\beta})$is its corresponding estimated variance. For every k, W_k has a limiting χ^2_1 distribution under H_0. The critical values are $\chi^2_{1,.05} = 3.84$ and $F^{.05}_{1,14} = 4.6$. The latter is an exact distribution test for $\beta = 1$ under H_0. Lafontaine and White (1985) try different integer exponents $(\pm k)$ where $k = 1, 2, 3, 6, 10, 20, 40$. Using $\widehat{\beta} = 1.14$ and $V(\widehat{\beta}) = (0.16)^2$ one gets $W_{-20} = 24.56$, $W_1 = 0.84$, and $W_{20} = 0.12$. The authors conclude that one could get any Wald statistic desired by choosing an appropriate exponent. Since $\beta > 1$, W_k is inversely related to k. So, we can find a W_k that exceeds the critical values given by the χ^2 and F distributions. In fact, W_{-20} leads to rejection whereas W_1 and W_{20} do not reject H_0.

For testing nonlinear restrictions, the Wald test is easy to compute. However, it has a serious problem in that it is not invariant to the way the null hypothesis is formulated. In this case, the score test may be difficult to compute, but Neyman's $C(\alpha)$ test is convenient to use and provide the invariance that is needed, see Dagenais and Dufour (1991).

Notes

1. For example, in a time-series setting, including the time trend in the multiple regression is equivalent to detrending each variable first, by residualing out the effect of time, and then running the regression on these residuals.

2. Two exceptions noted in Davidson and MacKinnon (1993) are the following: One, if the model is not identified asymptotically. For example, $y_t = \beta(1/t) + u_t$ for $t = 1, 2, \ldots, T$, will have $(1/t)$ tend to zero as $T \to \infty$. This means that as the sample size increase, there is no information on β. Two, if the number of parameters in the model increase as the sample size increase. For example, the fixed effects model in panel data discussed in Chapter 11.

3. If the MLE of β is $\widehat{\beta}_{MLE}$, then the MLE of $(1/\beta)$ is $(1/\widehat{\beta}_{MLE})$. Note that this invariance property implies that MLE cannot be in general unbiased. For example, even if $\widehat{\beta}_{MLE}$ is unbiased for β, by the above reparameterization, $(1/\widehat{\beta}_{MLE})$ is not unbiased for $(1/\beta)$.

4. If the distribution of disturbances is not normal, then OLS is still BLUE as long as the assumptions underlying the Gauss-Markov Theorem are satisfied. The MLE in this case will be in general more efficient than OLS as long as the distribution of the errors is correctly specified.

5. Using the Taylor Series approximation of $r(\widehat{\beta}_{MLE})$ around the true parameter vector β, one gets $r(\widehat{\beta}_{MLE}) \simeq r(\beta) + R(\beta)(\widehat{\beta}_{MLE} - \beta)$. Under the null hypothesis, $r(\beta) = 0$ and the $\text{var}[r(\widehat{\beta}_{MLE})] \simeq R(\beta)\ \text{var}(\widehat{\beta}_{MLE})R'(\beta)$.

Problems

1. *Invariance of the Fitted Values and Residuals to Nonsingular Transformations of the Independent Variables.* Post-multiply the independent variables in (7.1) by a nonsingular transformation C, so that $X^* = XC$.

   (a) Show that $P_{X^*} = P_X$ and $\bar{P}_{X^*} = \bar{P}_X$. Conclude that the regression of y on X has the same fitted values and the same residuals as the regression of y on X^*.

   (b) As an application of these results, suppose that every X was multiplied by a constant, say, a change in the units of measurement. Would the fitted values or residuals change when we rerun this regression?

   (c) Suppose that X contains two regressors X_1 and X_2 each of dimension $n \times 1$. If we run the regression of y on $(X_1 - X_2)$ and $(X_1 + X_2)$, will this yield the same fitted values and the same residuals as the original regression of y on X_1 and X_2?

2. *The FWL Theorem.*

   (a) Using partitioned inverse results from the Appendix, show that the solution to (7.9) yields $\widehat{\beta}_{2,OLS}$ given in (7.10).

(b) Alternatively, write (7.9) as a system of two equations in two unknowns $\widehat{\beta}_{1,OLS}$ and $\widehat{\beta}_{2,OLS}$. Solve, by eliminating $\widehat{\beta}_{1,OLS}$ and show that the resulting solution is given by (7.10).

(c) Using the FWL Theorem, show that if $X_1 = \iota_n$ a vector of ones indicating the presence of the constant in the regression, and X_2 is a set of economic variables, then (i) $\widehat{\beta}_{2,OLS}$ can be obtained by running $y_i - \bar{y}$ on the set of variables in X_2 expressed as deviations from their respective sample means. (ii) The least squares estimate of the constant $\widehat{\beta}_{1,OLS}$ can be retrieved as $\bar{y} - \bar{X}_2'\widehat{\beta}_{2,OLS}$ where $\bar{X}_2' = \iota_n' X_2/n$ is the vector of sample means of the independent variables in X_2.

3. Let $y = X\beta + D_i\gamma + u$ where y is $n \times 1$, X is $n \times k$ and D_i is a dummy variable that takes the value 1 for the i-th observation and 0 otherwise. Using the FWL Theorem, prove that the least squares estimates of β and γ from this regression are $\widehat{\beta}_{OLS} = (X^{*\prime}X^*)^{-1}X^{*\prime}y^*$ and $\widehat{\gamma}_{OLS} = y_i - x_i'\widehat{\beta}_{OLS}$, where X^* denotes the X matrix without the i-th observation and y^* is the y vector without the i-th observation and (y_i, x_i') denotes the i-th observation on the dependent and independent variables. This means that $\widehat{\gamma}_{OLS}$ is the forecasted OLS residual from the regression of y^* on X^* for the i-th observation which was essentially excluded from the regression by the inclusion of the dummy variable D_i.

4. *Maximum Likelihood Estimation.* Given the log-likelihood in (7.17),

(a) Derive the first-order conditions for maximization and show that $\widehat{\beta}_{MLE} = \widehat{\beta}_{OLS}$ and that $\widehat{\sigma}^2_{MLE} = RSS/n$.

(b) Calculate the second derivatives given in (7.18) and verify that the information matrix reduces to (7.19).

5. Given that $u \sim N(0, \sigma^2 I_n)$, we showed that $(n-k)s^2/\sigma^2 \sim \chi^2_{n-k}$. Use this fact to prove that,

(a) s^2 is unbiased for σ^2.

(b) $\text{var}(s^2) = 2\sigma^4/(n-k)$. **Hint:** $E(\chi^2_r) = r$ and $\text{var}(\chi^2_r) = 2r$.

6. Consider all estimators of σ^2 of the type $\widetilde{\sigma}^2 = e'e/r = u'\bar{P}_X u/r$ with $u \sim N(0, \sigma^2 I_n)$.

(a) Find $E(\widehat{\sigma}^2_{MLE})$ and the bias$(\widehat{\sigma}^2_{MLE})$.

(b) Find $\text{var}(\widehat{\sigma}^2_{MLE})$ and the MSE$(\widehat{\sigma}^2_{MLE})$.

(c) Compute MSE$(\widetilde{\sigma}^2)$ and minimize it with respect to r. Compare with the MSE of s^2 and $\widehat{\sigma}^2_{MLE}$.

7. *Computing Forecasts and Forecast Standard Errors Using a Regression Package.* This is based on Salkever (1976). From equations (7.23) and (7.24), show that

(a) $\widehat{\delta}'_{OLS} = (\widehat{\beta}'_{OLS}, \widehat{\gamma}'_{OLS})$ where $\widehat{\beta}'_{OLS} = (X'X)^{-1}X'y$, and $\widehat{\gamma}'_{OLS} = y_o - X_o\widehat{\beta}_{OLS}$. **Hint:** Set up the OLS normal equations and solve two equations in two unknowns. Alternatively, one can use the FWL Theorem to residual out the additional T_o dummy variables.

(b) $e^*_{OLS} = (e'_{OLS}, 0')'$ and $s^{*2} = s^2$.

(c) $s^{*2}(X^{*\prime}X^*)^{-1}$ is given by the expression in (7.25). **Hint:** Use partitioned inverse.

(d) Replace y_o by 0 and I_{T_o} by $-I_{T_o}$ in (7.23) and show that $\widehat{\gamma} = \widehat{y}_o = X_o\widehat{\beta}_{OLS}$ whereas all the results in parts (a), (b) and (c) remain the same.

8. (a) Show that $\text{cov}(\widehat{\beta}_{OLS}, e) = 0$. (Since both random variables are normally distributed, this proves their independence).

(b) Show that $\widehat{\beta}_{OLS}$ and s^2 are independent. **Hint:** A linear (Bu) and quadratic $(u'Au)$ forms in normal random variables are independent if $BA = 0$. See Graybill (1961) Theorem 4.17.

9. (a) Show that if one replaces R by c' in (7.29) one gets the square of the z-statistic given in (7.26).

(b) Show that when we replace σ^2 by s^2, the χ^2_1 statistic given in part (a) becomes the square of a t-statistic which is distributed as $F(1, n-K)$. **Hint:** The square of a $N(0,1)$ is χ^2_1. Also the ratio of two independent χ^2 random variables divided by their degrees of freedom is an F-statistic with these corresponding degrees of freedom, see Chapter 2.

10. (a) Show that the matrix A defined in (7.30) by $u'Au/\sigma^2$ is symmetric, idempotent and of rank g.

(b) Using the same proof given below lemma 1, show that (7.30) is χ^2_g.

11. (a) Show that the two quadratic forms $s^2 = u'\bar{P}_X u/(n-k)$ and that given in (7.30) are independent. **Hint:** Two positive semi-definite quadratic forms $u'Au$ and $u'Bu$ are independent if and only if $AB = 0$, see Graybill (1961) Theorem 4.10.

(b) Conclude that (7.31) is distributed as an $F(g, n-k)$.

12. *Restricted Least Squares.*

(a) Show that $\widehat{\beta}_{RLS}$ given by (7.36) is biased unless $R\beta = r$.

(b) Show that the $\text{var}(\widehat{\beta}_{RLS}) = \text{var}(A(X'X)^{-1}X'u)$ where

$$A = I_K - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R.$$

Prove that $A^2 = A$, but $A' \neq A$. Conclude that

$$\begin{aligned}\text{var}(\widehat{\beta}_{RLS}) = \quad & \sigma^2 A(X'X)^{-1}A' = \sigma^2\{(X'X)^{-1} \\ & -(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}\}.\end{aligned}$$

(c) Show that $\text{var}(\widehat{\beta}_{OLS}) - \text{var}(\widehat{\beta}_{RLS})$ is a positive semi-definite matrix.

13. *The Chow Test.*

(a) Show that OLS on (7.47) yields OLS on each equation separately in (7.46). In other words, $\widehat{\beta}_{1,OLS} = (X_1'X_1)^{-1}X_1'y_1$ and $\widehat{\beta}_{2,OLS} = (X_2'X_2)^{-1}X_2'y_2$.

(b) Show that the residual sum of squares for equation (7.47) is given by $RSS_1 + RSS_2$, where RSS_i is the residual sum of squares from running y_i on X_i for $i = 1, 2$.

(c) Show that the Chow F-statistic can be obtained from (7.49) by testing for the joint significance of H_o; $\beta_2 - \beta_1 = 0$.

14. Suppose we would like to test H_o; $\beta_2 = 0$ in the following unrestricted model given also in (7.8)

$$y = X\beta + u = X_1\beta_1 + X_2\beta_2 + u$$

(a) Using the FWL Theorem, show that the URSS is identical to the residual sum of squares obtained from $\bar{P}_{X_1}y = \bar{P}_{X_1}X_2\beta_2 + \bar{P}_{X_1}u$. Conclude that

$$URSS = y'\bar{P}_X y = y'\bar{P}_{X_1}y - y'\bar{P}_{X_1}X_2(X_2'\bar{P}_{X_1}X_2)^{-1}X_2'\bar{P}_{X_1}y.$$

(b) Show that the numerator of the F-statistic for testing H_o; $\beta_2 = 0$ which is given in (7.45), is $y'\bar{P}_{X_1}X_2(X_2'\bar{P}_{X_1}X_2)^{-1}X_2'\bar{P}_{X_1}y/k_2$.
Substituting $y = X_1\beta_1 + u$ under the null hypothesis, show that the above expression reduces to $u'\bar{P}_{X_1}X_2(X_2'\bar{P}_{X_1}X_2)^{-1}X_2'\bar{P}_{X_1}u/k_2$.

(c) Let $v = X_2'\bar{P}_{X_1}u$, show that if $u \sim \text{IIN}(0, \sigma^2)$ then $v \sim N(0, \sigma^2 X_2'\bar{P}_{X_1}X_2)$. Conclude that the numerator of the F-statistic given in part (b) when divided by σ^2 can be written as $v'[\text{var}(v)]^{-1}v/k_2$ where $v'[\text{var}(v)]^{-1}v$ is distributed as $\chi^2_{k_2}$ under H_o. **Hint**: See the discussion below lemma 1.

(d) Using the result that $(n-k)s^2/\sigma^2 \sim \chi^2_{n-k}$ where s^2 is the $URSS/(n-k)$, show that the F-statistic given by (7.45) is distributed as $F(k_2, n-k)$ under H_o. **Hint:** You need to show that $u'\bar{P}_X u$ is independent of the quadratic term given in part (b), see problem 11.

(e) Show that the Wald Test for H_o; $\beta_2 = 0$, given in (7.41), reduces in this case to $W = \widehat{\beta}_2'[R(X'X)^{-1}R']^{-1}\widehat{\beta}_2/s^2$ were $R = [0, I_{k_2}]$, $\widehat{\beta}_2$ denotes the OLS or equivalently the MLE of β_2 from the unrestricted model and s^2 is the corresponding estimate of σ^2 given by $URSS/(n-k)$. Using partitioned inversion or the FWL Theorem, show that the numerator of W is k_2 times the expression in part (b).

(f) Show that the score form of the LM statistic, given in (7.42) and (7.44), can be obtained as the explained sum of squares from the artificial regression of the restricted residuals $(y - X_1\widehat{\beta}_{1,RLS})$ deflated by $\widetilde{s}$ on the matrix of regressors X. In this case, $\widetilde{s}^2 = RRSS/(n-k_1)$ is the Mean Square Error of the restricted regression. In other words, obtain the explained sum of squares from regressing $\bar{P}_{X_1}y/\widetilde{s}$ on X_1 and X_2.

15. *Iterative Estimation in Partitioned Regression Models.* This is based on Fiebig (1995). Consider the partitioned regression model given in (7.8) and let X_2 be a single regressor, call it x_2 of dimension $n \times 1$ so that β_2 is a scalar. Consider the following strategy for estimating β_2: Estimate β_1 from the shortened regression of y on X_1. Regress the residuals from this regression on x_2 to yield $b_2^{(1)}$.

(a) Prove that $b_2^{(1)}$ is biased.

Now consider the following iterative strategy for re-estimating β_2:

Re-estimate β_1 by regressing $y - x_2 b_2^{(1)}$ on X_1 to yield $b_1^{(1)}$. Next iterate according to the following scheme:

$$b_1^{(j)} = (X_1'X_1)^{-1}X_1'(y - x_2 b_2^{(j)})$$
$$b_2^{(j+1)} = (x_2'x_2)^{-1}x_2'(y - X_1 b_1^{(j)}), \quad j = 1, 2, ...$$

(b) Determine the behavior of the bias of $b_2^{(j+1)}$ as j increases.

(c) Show that as j increases $b_2^{(j+1)}$ converges to the estimator of β_2 obtained by running OLS on (7.8).

16. *Maddala (1992, pp. 120-127).* Consider the simple linear regression

$$y_i = \alpha + \beta X_i + u_i \quad i = 1, 2, \ldots, n.$$

where α and β are scalars and $u_i \sim \text{IIN}(0, \sigma^2)$. For H_o; $\beta = 0$,

(a) Derive the Likelihood Ratio (LR) statistic and show that it can be written as $n\log[1/(1-r^2)]$ where r^2 is the square of the correlation coefficient between X and y.

(b) Derive the Wald (W) statistic for testing H_o; $\beta = 0$. Show that it can be written as $nr^2/(1-r^2)$. This is the square of the usual t-statistic on β with $\widehat{\sigma}^2_{MLE} = \sum_{i=1}^n e_i^2/n$ used instead of s^2 in estimating σ^2. $\widehat{\beta}$ is the unrestricted MLE which is OLS in this case, and the e_i's are the usual least squares residuals.

(c) Derive the Lagrange Multiplier (LM) statistic for testing H_o; $\beta = 0$. Show that it can be written as nr^2. This is the square of the usual t-statistic on with $\widetilde{\sigma}^2_{RMLE} = \sum_{i=1}^n (y_i - \bar{y})^2/n$ used instead of s^2 in estimating σ^2. The $\widetilde{\sigma}^2_{RMLE}$ is restricted MLE of σ^2 (i.e., imposing H_o and maximizing the likelihood with respect to σ^2).

(d) Show that $LM/n = (W/n)/[1 + (W/n)]$, and $LR/n = \log[1 + (W/n)]$. Using the following inequality $x \geq \log(1 + x) \geq x/(1 + x)$, conclude that $W \geq LR \geq LM$. **Hint:** Use $x = W/n$.

(e) For the cigarette consumption data given in Table 3.2, compute the W, LR, LM for the simple regression of log$C$ on log$P$ and demonstrate the above inequality given in part (d) for testing that the price elasticity is zero?

17. *Engle (1984, pp.785-786).* Consider a set of T independent observations on a Bernoulli random variable which takes on the values $y_t = 1$ with probability θ, and $y_t = 0$ with probability $(1 - \theta)$.

(a) Derive the log-likelihood function, the MLE of θ, the score $S(\theta)$, and the information $I(\theta)$.

(b) Compute the LR, W and LM test statistics for testing H_o; $\theta = \theta_o$, versus H_A; $\theta \neq \theta_o$ for $\theta \epsilon (0, 1)$.

18. *Engle (1984, pp. 787-788).* Consider the linear regression model

$$y = X\beta + u = X_1\beta_1 + X_2\beta_2 + u$$

given in (7.8), where $u \sim N(0, \sigma^2 I_T)$.

(a) Write down the log-likelihood function, find the MLE of β and σ^2.

(b) Write down the score $S(\beta)$ and show that the information matrix is block-diagonal between β and σ^2.

(c) Derive the W, LR and LM test statistics in order to test H_o; $\beta_1 = \beta_1^o$, versus H_A; $\beta_1 \neq \beta_1^o$, where β_1 is say the first k_1 elements of β. Show that if $X = [X_1, X_2]$, then

$$W = (\beta_1^o - \widehat{\beta}_1)'[X_1'\bar{P}_{X_2}X_1](\beta_1^o - \widehat{\beta}_1)/\widehat{\sigma}^2$$
$$LM = \widetilde{u}'X_1[X_1'\bar{P}_{X_2}X_1]^{-1}X_1'\widetilde{u}/\widetilde{\sigma}^2$$
$$LR = T\ \log(\widetilde{u}'\widetilde{u}/\widehat{u}'\widehat{u})$$

where $\widehat{u} = y - X\widehat{\beta}$, $\widetilde{u} = y - X\widetilde{\beta}$ and $\widehat{\sigma}^2 = \widehat{u}'\widehat{u}/T$, $\widetilde{\sigma}^2 = \widetilde{u}'\widetilde{u}/T$. $\widehat{\beta}$ is the unrestricted MLE, whereas $\widetilde{\beta}$ is the restricted MLE.

(d) Using the above results, show that

$$W = T(\widetilde{u}'\widetilde{u} - \widehat{u}'\widehat{u})/\widehat{u}'\widehat{u}$$
$$LM = T(\widetilde{u}'\widetilde{u} - \widehat{u}'\widehat{u})/\widetilde{u}'\widetilde{u}$$

Also, that $LR = T\ \log[1 + (W/T)]$; $LM = W/[1 + (W/T)]$; and $(T - k)W/Tk_1 \sim F_{k_1,T-k}$ under H_o. As in problem 16, we use the inequality $x \geq \log(1 + x) \geq x/(1 + x)$ to conclude that $W \geq LR \geq LM$. **Hint:** Use $x = W/T$. However, it is important to note that all the test statistics are monotonic functions of the F-statistic and exact tests for each would produce identical critical regions.

(e) For the cigarette consumption data given in Table 3.2, run the following regression:

$$\log C = \alpha + \beta \log P + \gamma \log Y + u$$

compute the W, LR, LM given in part (c) for the null hypothesis H_o; $\beta = -1$.

(f) Compute the Wald statistics for H_o^A; $\beta = -1$, H_o^B; $\beta^5 = -1$ and H_o^C; $\beta^{-5} = -1$. How do these statistics compare?

19. *Gregory and Veall (1985).* Using equation (7.51) and the two formulations of the null hypothesis H^A and H^B given below (7.50), verify that the Wald statistics corresponding to these two formulations are those given in (7.52) and (7.53), respectively.

20. *Gregory and Veall (1986).* Consider the dynamic equation

$$y_t = \rho y_{t-1} + \beta_1 x_t + \beta_2 x_{t-1} + u_t$$

where $|\rho| < 1$, and $u_t \sim \text{NID}(0, \sigma^2)$. Note that for this equation to be the Cochrane-Orcutt transformation

$$y_t - \rho y_{t-1} = \beta_1(x_t - \rho x_{t-1}) + u_t$$

the following nonlinear restriction must be satisfied $-\beta_1\rho = \beta_2$ called the *common factor* restriction by Hendry and Mizon (1978). Now consider the following four formulations of this restriction H^A; $\beta_1\rho + \beta_2 = 0$; H^B; $\beta_1 + (\beta_2/\rho) = 0$; H^C; $\rho + (\beta_2/\beta_1) = 0$ and H^D; $(\beta_1\rho/\beta_2) + 1 = 0$.

   (a) Using equation (7.51) derive the four Wald statistics corresponding to the four formulations of the null hypothesis.

   (b) Apply these four Wald statistics to the equation relating real personal consumption expenditures to real disposable personal income in the U.S. over the post World War II period 1950-1993, see Table 5.1.

21. *Effect of Additional Regressors on R^2.* This problem was considered in non-matrix form in Chapter 4, problem 4. Regress y on X_1 which is $T \times K_1$ and compute SSE_1. Add X_2 which is $T \times K_2$ so that the number of regressors in now $K = K_1 + K_2$. Regress y on $X = [X_1, X_2]$ and get SSE_2. Show that $SSE_2 \leq SSE_1$. Conclude that the corresponding R-squares satisfy $R_2^2 \geq R_1^2$. **Hint:** Show that $P_X - P_{X_1}$ is a positive semi-definite matrix.

References

Additional readings for the material covered in this chapter can be found in Davidson and MacKinnon (1993), Kelejian and Oates (1989), Maddala (1992), Fomby, Hill and Johnson (1984), Greene (1993), Johnston (1984), Judge et al. (1985) and Theil (1971). These econometrics texts were cited earlier. Other references cited in this chapter are the following:

Bera A.K. and G. Permaratne (2001), "General Hypothesis Testing," Chapter 2 in Baltagi, B.H. (ed.), *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Berndt, E.R. and N.E. Savin (1977), "Conflict Among Criteria for Testing Hypotheses in the Multivariate Linear Regression Model," *Econometrica*, 45: 1263-1278.

Buse, A.(1982), "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note," *The American Statistician*, 36: 153-157.

Chow, G.C. (1960), "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, 28: 591-605.

Dagenais, M.G. and J. M. Dufour (1991), "Invariance, Nonlinear Models, and Asymptotic Tests," *Econometrica*, 59: 1601-1615.

Engle, R.F. (1984), "Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics," In: Griliches, Z. and M.D. Intrilligator (eds) *Handbook of Econometrics* (North-Holland: Amsterdam).

Fiebig, D.G. (1995), "Iterative Estimation in Partitioned Regression Models," *Econometric Theory*, Problem 95.5.1, 11:1177.

Frisch, R., and F.V. Waugh (1933), "Partial Time Regression as Compared with Individual Trends," *Econometrica*, 1: 387-401.

Graybill, F.A.(1961), *An Introduction to Linear Statistical Models*, Vol. 1 (McGraw-Hill: New York).

Gregory, A.W. and M.R. Veall (1985), "Formulating Wald Tests of Nonlinear Restrictions," *Econometrica*, 53: 1465-1468.

Gregory, A.W. and M.R. Veall (1986), "Wald Tests of Common Factor Restrictions," *Economics Letters*, 22: 203-208.

Gujarati, D. (1970), "Use of Dummy Variables in Testing for Equality Between Sets of Coefficients in Two Linear Regressions: A Generalization," *The American Statistician*, 24: 50-52.

Hendry, D.F. and G.E. Mizon (1978), "Serial Correlation as a Convenient Simplification, Not as a Nuisance: A Comment on A Study of the Demand for Money by the Bank of England," *Economic Journal*, 88: 549-563.

Lafontaine, F. and K.J. White (1986), "Obtaining Any Wald Statistic You Want," *Economics Letters*, 21: 35-40.

Lovell, M.C. (1963), "Seasonal Adjustment of Economic Time Series," *Journal of the American Statistical Association*, 58: 993-1010.

Salkever, D. (1976), "The Use of Dummy Variables to Compute Predictions, Prediction Errors, and Confidence Intervals," *Journal of Econometrics*, 4: 393-397.

Appendix
Some Useful Matrix Properties

This book assumes that the reader has encountered matrices before, and knows how to add, subtract and multiply conformable matrices. In addition, that the reader is familiar with the transpose, trace, rank, determinant and inverse of a matrix. Unfamiliar readers should consult standard texts like Bellman (1970) or Searle (1982). The purpose of this Appendix is to review some useful matrix properties that are used in the text and provide easy access to these properties. Most of these properties are given without proof.

Starting with Chapter 7, our data matrix X is organized such that it has n rows and k columns, so that each row denotes an observation on k variables and each column denotes n observations on one variable. This matrix is of dimension $n \times k$. The rank of an $n \times k$ matrix is always less than or equal to its smaller dimension. Since $n > k$, the rank $(X) \leq k$. When there is no perfect multicollinearity among the variables in X, this matrix is said to be of full column rank k. In this case, $X'X$, the matrix of cross-products is of dimension $k \times k$. It is square, symmetric and of full rank k. This uses the fact that the rank$(X'X)$ = rank$(X) = k$. Therefore, $(X'X)$ is nonsingular and the inverse $(X'X)^{-1}$ exists. This is needed for the computation of *Ordinary Least Squares*. In fact, for least squares to be feasible, X should be of full column rank k and no variable in X should be a perfect linear combination of the other variables in X. If we write

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}$$

where x_i' denotes the i-th observation, in the data, then $X'X = \sum_{i=1}^{n} x_i x_i'$ where x_i is a column vector of dimension $k \times 1$.

An important and widely encountered matrix is the *Identity matrix* which will be denoted by I_n and subscripted by its dimension n. This is a square $n \times n$ matrix whose diagonal elements are all equal to one and its off diagonal elements are all equal to zero. Also, $\sigma^2 I_n$ will be a familiar *scalar covariance matrix*, with every diagonal element equal to σ^2 reflecting *homoskedasticity* or equal variances (see Chapter 5), and zero covariances or no *serial correlation* (see Chapter 5). Let

$$\Omega = \text{diag}[\sigma_i^2] = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{bmatrix}$$

be an $(n \times n)$ *diagonal* matrix with the i-th diagonal element equal to σ_i^2 for $i = 1, 2, \ldots, n$. This matrix will be encountered under heteroskedasticity, see Chapter 9. Note that $\text{tr}(\Omega) = \sum_{i=1}^{n} \sigma_i^2$ is the sum of its diagonal elements. Also, $\text{tr}(I_n) = n$ and $\text{tr}(\sigma^2 I_n) = n\sigma^2$. Another useful matrix is the *projection matrix* $P_X = X(X'X)^{-1}X'$ which is of dimension $n \times n$. This matrix is encountered in Chapter 7. If y denotes the $n \times 1$ vector of observations on the dependent variable, then $P_X y$ generates the predicted values $\widehat{y}$ from the least squares regression of y on X. This matrix P_X is symmetric and *idempotent*. This means that $P_X' = P_X$ and $P_X^2 = P_X P_X = P_X$ as can be easily verified. Some of the properties of idempotent matrices is that their rank is equal to their trace. Hence, $\text{rank}(P_X) = \text{tr}(P_X) = \text{tr}[X(X'X)^{-1}X'] = \text{tr}[X'X(X'X^{-1})] = \text{tr}(I_k) = k$.

Here, we used the fact that $\text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA)$. In other words, the trace is unaffected by the cyclical permutation of the product. Of course, these matrices should be conformable and the product should result in a square matrix. Note that $\bar{P}_X = I_n - P_X$ is also a symmetric and idempotent matrix. In this case, $\bar{P}_X y = y - P_X y = y - \widehat{y} = e$ where e denotes the least squares residuals, $y - X\widehat{\beta}_{OLS}$ where $\widehat{\beta}_{OLS} = (X'X)^{-1}X'y$, see Chapter 7. Some properties of these projection matrices are the following:

$$P_X X = X, \bar{P}_X X = 0, \bar{P}_X e = e \quad \text{and} \quad P_X e = 0.$$

In fact, $X'e = 0$ means that the matrix X is *orthogonal* to the vector of least squares residuals e. Note that $X'e = 0$ means that $X'(y - X\widehat{\beta}_{OLS}) = 0$ or $X'y = X'X\widehat{\beta}_{OLS}$. These k equations are known as the OLS normal equations and their solution yields the least squares estimates $\widehat{\beta}_{OLS}$. By the definition of $\bar{P}_X$, we have (i) $P_X + \bar{P}_X = I_n$. Also, (ii) P_X and $\bar{P}_X$ are idempotent and (iii) $P_X \bar{P}_X = 0$. In fact, any two of these properties imply the third. The $\text{rank}(\bar{P}_X) = \text{tr}(\bar{P}_X) = \text{tr}(I_n - P_X) = n - k$. Note that P_X and $\bar{P}_X$ are of rank k and $(n - k)$, respectively. Both matrices are not of full column rank. In fact, the only full rank, symmetric idempotent matrix is the identity matrix.

Matrices not of full rank are singular, and their inverse do not exist. However, one can find a *generalized inverse* of a matrix Ω which we will call Ω^- which satisfies the following requirements:

(i) $\Omega\Omega^-\Omega = \Omega$ (ii) $\Omega^-\Omega\Omega^- = \Omega^-$
(iii) $\Omega^-\Omega$ is symmetric and (iv) $\Omega\Omega^-$ is symmetric.

Even if Ω is not square, a unique Ω^- can be found for Ω which satisfies the above four properties. This is called the *Moore-Penrose* generalized inverse.

Note that a symmetric idempotent matrix is its own Moore-Penrose generalized inverse. For example, it is easy to verify that if $\Omega = P_X$, then $\Omega^- = P_X$ and that it satisfies the above four properties. Idempotent matrices have *characteristic roots* that are either zero or one. The number of non-zero characteristic roots is equal to the rank of this matrix. The characteristic roots of Ω^{-1} are the reciprocals of the characteristic roots of Ω, but the characteristic vectors of both matrices are the same.

The determinant of a matrix is non-zero if and only if it has full rank. Therefore, if A is singular, then $|A| = 0$. Also, the determinant of a matrix is equal to the product of its characteristic roots. For two square matrices A and B, the determinant of the product is the product of the determinants $|AB| = |A| \cdot |B|$. Therefore, the determinant of Ω^{-1} is the reciprocal of the determinant of Ω. This follows from the fact that $|\Omega||\Omega^{-1}| = |\Omega\Omega^{-1}| = |I| = 1$. This property is used in writing the likelihood function for *Generalized Least Squares* (GLS) estimation, see Chapter 9. The determinant of a triangular matrix

is equal to the product of its diagonal elements. Of course, it immediately follows that the determinant of a diagonal matrix is the product of its diagonal elements.

The constant in the regression corresponds to a vector of ones in the matrix of regressors X. This vector of ones is denoted by ι_n where n is the dimension of this column vector. Note that $\iota_n'\iota_n = n$ and $\iota_n\iota_n' = J_n$ where J_n is a matrix of ones of dimension $n \times n$. Note also that J_n is not idempotent, but $\bar{J}_n = J_n/n$ is idempotent as can be easily verified. The $\text{rank}(\bar{J}_n) = \text{tr}(\bar{J}_n) = 1$. Note also that $I_n - \bar{J}_n$ is idempotent with rank $(n-1)$. $\bar{J}_n y$ has a typical element $\bar{y} = \sum_{i=1}^{n} y_i/n$ whereas $(I_n - \bar{J}_n)y$ has a typical element $(y_i - \bar{y})$. So that $\bar{J}_n$ is the *averaging* matrix, whereas premultiplying by $(I_n - \bar{J}_n)$ results in deviations from the mean.

For two nonsingular matrices A and B

$$(AB)^{-1} = B^{-1}A^{-1}$$

Also, the transpose of a product of two conformable matrices, $(AB)' = B'A'$. In fact, for the product of three conformable matrices this becomes $(ABC)' = C'B'A'$. The transpose of the inverse is the inverse of the transpose, i.e., $(A^{-1})' = (A')^{-1}$.

The inverse of a partitioned matrix

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

is

$$A^{-1} = \begin{bmatrix} E & -EA_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}E & A_{22}^{-1} + A_{22}^{-1}A_{21}EA_{12}A_{22}^{-1} \end{bmatrix}$$

where $E = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}$. Alternatively, it can be expressed as

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}FA_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}F \\ -FA_{21}A_{11}^{-1} & F \end{bmatrix}$$

where $F = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$. These formulas are used in partitioned regression models, see for example the Frisch-Waugh Lovell Theorem and the computation of the variance-covariance matrix of forecasts from a multiple regression in Chapter 7.

An $n \times n$ symmetric matrix Ω has n distinct characteristic vectors $c_1, \ldots, c_n$. The corresponding n characteristic roots $\lambda_1, \ldots, \lambda_n$ may not be distinct but they are all real numbers. The number of non-zero characteristic roots of Ω is equal to the rank of Ω. The characteristic roots of a positive definite matrix are positive. The characteristic vectors of the symmetric matrix Ω are *orthogonal* to each other, i.e., $c_i'c_j = 0$ for $i \neq j$ and can be made *orthonormal* with $c_i'c_i = 1$ for $i = 1, 2, \ldots, n$. Hence, the matrix of characteristic vectors $C = [c_1, c_2, \ldots, c_n]$ is an *orthogonal* matrix, such that $CC' = C'C = I_n$ with $C' = C^{-1}$. By definition $\Omega c_i = \lambda_i c_i$ or $\Omega C = C\Lambda$ where $\Lambda = \text{diag}[\lambda_i]$. Premultiplying the last equation by C' we get $C'\Omega C = C'C\Lambda = \Lambda$. Therefore, the matrix of characteristic vectors C diagonalizes the symmetric matrix Ω. Alternatively, we can write $\Omega = C\Lambda C' = \sum_{i=1}^{n} \lambda_i c_i c_i'$ which is the *spectral decomposition* of Ω.

A real symmetric $n \times n$ matrix Ω is *positive semi-definite* if for every $n \times 1$ non-negative vector y, we have $y'\Omega y \geq 0$. If $y'\Omega y$ is strictly positive for any non-zero y then Ω is said to be *positive definite*. A necessary and sufficient condition for Ω to be positive definite is that all the characteristic roots of Ω are positive. One important application is the comparison of efficiency of two unbiased estimators of a vector of parameters β. In this case, we subtract the variance-covariance matrix of the inefficient estimator from the more efficient one and show that the resulting difference yields a positive semi-definite matrix, see the Gauss-Markov Theorem in Chapter 7.

If Ω is a symmetric and positive definite matrix, there exists a nonsingular matrix P such that $\Omega = PP'$. In fact, using the spectral decomposition of Ω given above, one choice for $P = C\Lambda^{1/2}$ so

that $\Omega = C\Lambda C' = PP'$. This is a useful result which we use in Chapter 9 to obtain Generalized Least Squares (GLS) as a least squares regression after transforming the original regression model by $P^{-1} = (C\Lambda^{1/2})^{-1} = \Lambda^{-1/2}C'$. In fact, if $u \sim (0, \sigma^2\Omega)$, then $P^{-1}u$ has zero mean and $\text{var}(P^{-1}u) = P^{-1\prime}\text{var}(u)P^{-1\prime} = \sigma^2 P^{-1}\Omega P^{-1\prime} = \sigma^2 P^{-1}PP'P^{-1\prime} = \sigma^2 I_n$.

From Chapter 2, we have seen that if $u \sim N(0, \sigma^2 I_n)$, then $u_i/\sigma \sim N(0,1)$, so that $u_i^2/\sigma^2 \sim \chi_1^2$ and $u'u/\sigma^2 = \sum_{i=1}^n u_i^2/\sigma^2 \sim \chi_n^2$. Therefore, $u'(\sigma^2 I_n)^{-1}u \sim \chi_n^2$. If $u \sim N(0, \sigma^2\Omega)$ where Ω is positive definite, then $u^* = P^{-1}u \sim N(0, \sigma^2 I_n)$ and $u^{*\prime}u^*/\sigma^2 \sim \chi_n^2$. But $u^{*\prime}u^* = u'P^{-1\prime}P^{-1}u = u'\Omega^{-1}u$. Hence, $u'\Omega^{-1}u/\sigma^2 \sim \chi_n^2$. This is used in Chapter 9.

Note that the OLS residuals are denoted by $e = \bar{P}_X u$. If $u \sim N(0, \sigma^2 I_n)$, then e has mean zero and $\text{var}(e) = \sigma^2 \bar{P}_X I_n \bar{P}_X = \sigma^2 \bar{P}_X$ so that $e \sim N(0, \sigma^2 \bar{P}_X)$. Our estimator of σ^2 in Chapter 7 is $s^2 = e'e/(n-k)$ so that $(n-k)s^2/\sigma^2 = e'e/\sigma^2$. The last term can also be written as $u'\bar{P}_X u/\sigma^2$. In order to find the distribution of this quadratic form in Normal variables, we use the following result stated as lemma 1 in Chapter 7.

Lemma 1: For every symmetric idempotent matrix A of rank r, there exists an orthogonal matrix P such that $P'AP = J_r$ where J_r is a diagonal matrix with the first r elements equal to one and the rest equal to zero.

We use this lemma to show that the $e'e/\sigma^2$ is a chi-squared with $(n-k)$ degrees of freedom. To see this note that $e'e/\sigma^2 = u'\bar{P}_X u/\sigma^2$ and that $\bar{P}_X$ is symmetric and idempotent of rank $(n-k)$. Using the lemma there exists a matrix P such that $P'\bar{P}_X P = J_{n-k}$ is a diagonal matrix with the first $(n-k)$ elements on the diagonal equal to 1 and the last k elements equal to zero. An orthogonal matrix P is by definition a matrix whose inverse, is its own transpose, i.e., $P'P = I_n$. Let $v = P'u$ then v has mean zero and $\text{var}(v) = \sigma^2 P'P = \sigma^2 I_n$ so that v is $N(0, \sigma^2 I_n)$ and $u = Pv$. Therefore,

$$e'e/\sigma^2 = u'\bar{P}_X u/\sigma^2 = v'P'\bar{P}_X Pv/\sigma^2 = v'J_{n-k}v/\sigma^2 = \sum_{i=1}^{n-k} v_i^2/\sigma^2$$

But, the v's are independent identically distributed $N(0, \sigma^2)$, hence v_i^2/σ^2 is the square of a standardized $N(0,1)$ random variable which is distributed as a χ_1^2. Moreover, the sum of independent χ^2 random variables is a χ^2 random variable with degrees of freedom equal to the sum of the respective degrees of freedom, see Chapter 2. Hence, $e'e/\sigma^2$ is distributed as χ_{n-k}^2.

The beauty of the above result is that it applies to all quadratic forms $u'Au$ where A is symmetric and idempotent. In general, for $u \sim N(0, \sigma^2 I)$, a necessary and sufficient condition for $u'Au/\sigma^2$ to be distributed χ_k^2 is that A is idempotent of rank k, see Theorem 4.6 of Graybill (1961). Another useful theorem on quadratic forms in normal random variables is the following: If $u \sim N(0, \sigma^2\Omega)$, then $u'Au/\sigma^2$ is χ_k^2 if and only if $A\Omega$ is an idempotent matrix of rank k, see Theorem 4.8 of Graybill (1961). If $u \sim N(0, \sigma^2 I)$, the two positive semi-definite quadratic forms in normal random variables say $u'Au$ and $u'Bu$ are independent if and only if $AB = 0$, see Theorem 4.10 of Graybill (1961). A sufficient condition is that $\text{tr}(AB) = 0$, see Theorem 4.15 of Graybill (1961). This is used in Chapter 7 to construct F-statistics to test hypotheses, see for example problem 11. For $u \sim N(0, \sigma^2 I)$, the quadratic form $u'Au$ is independent of the linear form Bu if $BA = 0$, see Theorem 4.17 of Graybill (1961). This is used in Chapter 7 to prove the independence of s^2 and $\widehat{\beta}_{ols}$, see problem 8. In general, if $u \sim N(0, \Sigma)$, then $u'Au$ and $u'Bu$ are independent if and only if $A\Sigma B = 0$, see Theorem 4.21 of Graybill (1961). Many other useful matrix properties can be found. This is only a sample of them that will be implicitly or explicitly used in this book.

The *Kronecker* product of two matrices say $\Sigma \otimes I_n$ where Σ is $m \times m$ and I_n is the identity matrix of dimension n is defined as follows:

$$\Sigma \otimes I_n = \begin{bmatrix} \sigma_{11}I_n & \dots & \sigma_{1m}I_n \\ \vdots & & \vdots \\ \sigma_{m1}I_n & \dots & \sigma_{mm}I_n \end{bmatrix}$$

In other words, we place an I_n next to every element of $\Sigma = [\sigma_{ij}]$. The dimension of the resulting matrix is $mn \times mn$. This is useful when we have a system of equations like Seemingly Unrelated Regressions in Chapter 10. In general, if A is $m \times n$ and B is $p \times q$ then $A \otimes B$ is $mp \times nq$. Some properties of Kronecker

products include $(A \otimes B)' = A' \otimes B'$. If both A and B are square matrices of order $m \times m$ and $p \times p$ then $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}, |A \otimes B| = |A|^m |B|^p$ and $\text{tr}(A \otimes B) = \text{tr}(A)\text{tr}(B)$. Applying this result to $\Sigma \otimes I_n$ we get

$$(\Sigma \otimes I_n)^{-1} = \Sigma^{-1} \otimes I_n \quad \text{and} \quad |\Sigma \otimes I_n| = |\Sigma|^m |I_n|^n = |\Sigma|^m$$

and $\text{tr}(\Sigma \otimes I_n) = \text{tr}(\Sigma)\text{tr}(I_n) = n \ \text{tr}(\Sigma)$.

Some useful properties of matrix differentiation are the following:

$$\frac{\partial x'b}{\partial b} = x \quad \text{where } x' \text{ is } 1 \times k \text{ and } b \text{ is } k \times 1.$$

Also

$$\frac{\partial b'Ab}{\partial b} = (A + A') \quad \text{where } A \text{ is } k \times k.$$

If A is symmetric, then $\partial b'Ab/\partial b = 2Ab$. These two properties will be used in Chapter 7 in deriving the least squares estimator.

References

Bellman, R. (1970), *Introduction to Matrix Analysis* (McGraw Hill: New York).

Searle, S.R. (1982), *Matrix Algebra Useful for Statistics* (John Wiley and Sons: New York).

CHAPTER 8
Regression Diagnostics and Specification Tests

8.1 Influential Observations[1]

Sources of influential observations include: (i) improperly recorded data, (ii) observational errors in the data, (iii) misspecification and (iv) outlying data points that are legitimate and contain valuable information which improve the efficiency of the estimation. It is constructive to isolate extreme points and to determine the extent to which the parameter estimates depend upon these desirable data.

One should always run descriptive statistics on the data, see Chapter 2. This will often reveal outliers, skewness or multimodal distributions. Scatter diagrams should also be examined, but these diagnostics are only the first line of attack and are inadequate in detecting multivariate discrepant observations or the way each observation affects the estimated regression model.

In regression analysis, we emphasize the importance of plotting the residuals against the explanatory variables or the predicted values $\widehat{y}$ to identify patterns in these residuals that may indicate nonlinearity, heteroskedasticity, serial correlation, etc, see Chapter 3. In this section, we learn how to identify significantly large residuals and compute regression diagnostics that may identify influential observations. We study the extent to which the deletion of any observation affects the estimated coefficients, the standard errors, predicted values, residuals and test statistics. These represent the core of diagnostic tools in regression analysis.

Accordingly, Belsley, Kuh and Welsch (1980, p.11) define an influential observation as "..one which, either individually or together with several other observations, has demonstrably larger impact on the calculated values of various estimates (coefficients, standard errors, t-values, etc.) than is the case for most of the other observations."

First, what is a significantly large residual? We have seen that the least squares residuals of y on X are given by $e = (I_n - P_X)u$, see equation (7.7). y is $n \times 1$ and X is $n \times k$. If $u \sim \text{IID}(0, \sigma^2 I_n)$, then e has zero mean and variance $\sigma^2(I_n - P_X)$. Therefore, the OLS residuals are *correlated* and *heteroskedastic* with $\text{var}(e_i) = \sigma^2(1 - h_{ii})$ where h_{ii} is the i-th diagonal element of the *hat* matrix $H = P_X$, since $\widehat{y} = Hy$.

The diagonal elements h_{ii} have the following properties:

$$\sum_{i=1}^{n} h_{ii} = \text{tr}(P_X) = k \quad \text{and} \quad h_{ii} = \sum_{j=1}^{n} h_{ij}^2 \geq h_{ii}^2 \geq 0.$$

The last property follows from the fact that P_X is symmetric and idempotent. Therefore, $h_{ii}^2 - h_{ii} \leq 0$ or $h_{ii}(h_{ii} - 1) \leq 0$. Hence, $0 \leq h_{ii} \leq 1$, (see problem 1). h_{ii} is called the *leverage* of the i-th observation. For a simple regression with a constant,

$$h_{ii} = (1/n) + (x_i^2 / \sum_{i=1}^{n} x_i^2)$$

where $x_i = X_i - \bar{X}$; h_{ii} can be interpreted as a measure of the distance between X values of the i-th observation and their mean over all n observations. A large h_{ii} indicates that the i-th observation is distant from the center of the observations. This means that the i-th observation with large h_{ii} (a function only of X_i values) exercises substantial leverage in determining the fitted value $\widehat{y}_i$. Also, the larger h_{ii}, the smaller the variance of the residual e_i. Since observations

with high leverage tend to have smaller residuals, it may not be possible to detect them by an examination of the residuals alone. But, what is a large leverage? h_{ii} is large if it is more than twice the mean leverage value $2\bar{h} = 2k/n$. Hence, $h_{ii} \geq 2k/n$ are considered outlying observations with regards to X values.

An alternative representation of h_{ii} is simply $h_{ii} = d_i' P_X d_i = ||P_X d_i||^2 = x_i'(X'X)^{-1}x_i$ where d_i denotes the i-th observation's dummy variable, i.e., a vector of dimension n with 1 in the i-th position and 0 elsewhere. x_i' is the i-th row of X and $||.||$ denotes the Euclidian length. Note that $d_i'X = x_i'$.

Let us standardize the i-th OLS residual by dividing it by an estimate of its variance. A *standardized residual* would then be:

$$\widetilde{e}_i = e_i / s\sqrt{1 - h_{ii}} \tag{8.1}$$

where σ^2 is estimated by s^2, the MSE of the regression. This is an *internal studentization* of the residuals, see Cook and Weisberg (1982). Alternatively, one could use an estimate of σ^2 that is independent of e_i. Defining $s^2_{(i)}$ as the MSE from the regression computed without the i-th observation, it can be shown, see equation (8.18) below, that

$$s^2_{(i)} = \frac{(n-k)s^2 - e_i^2/(1-h_{ii})}{(n-k-1)} = s^2\left(\frac{n-k-\widetilde{e}_i^2}{n-k-1}\right) \tag{8.2}$$

Under normality, $s^2_{(i)}$ and e_i are independent and the *externally studentized* residuals are defined by

$$e_i^* = e_i / s_{(i)}\sqrt{1 - h_{ii}} \sim t_{n-k-1} \tag{8.3}$$

Thus, if the normality assumption holds, we can readily assess the significance of any single studentized residual. Of course, the e_i^* will not be independent. Since this is a t-statistic, it is natural to think of e_i^* as large if its value exceeds 2 in absolute value.

Substituting (8.2) into (8.3) and comparing the result with (8.1), it is easy to show that e_i^* is a monotonic transformation of $\widetilde{e}_i$

$$e_i^* = \widetilde{e}_i \left(\frac{n-k-1}{n-k-\widetilde{e}_i^2}\right)^{\frac{1}{2}} \tag{8.4}$$

Cook and Wiesberg (1982) show that e_i^* can be obtained as a t-statistic from the following augmented regression:

$$y = X\beta^* + d_i\varphi + u \tag{8.5}$$

where d_i is the dummy variable for the i-th observation. In fact, $\widehat{\varphi} = e_i/(1-h_{ii})$ and e_i^* is the t-statistic for testing that $\varphi = 0$. (see problem 4 and the proof given below). Hence, whether the i-th residual is large can be simply determined by the regression (8.5). A dummy variable for the i-th observation is included in the original regression and the t-statistic on this dummy tests whether this i-th residual is large. This is repeated for all observations $i = 1, \ldots, n$.

This can be generalized easily to testing for a group of significantly large residuals:

$$y = X\beta^* + D_p\varphi^* + u \tag{8.6}$$

where D_p is an $n \times p$ matrix of dummy variables for the p-suspected observations. One can test $\varphi^* = 0$ using the Chow test described in (4.17) as follows:

$$F = \frac{[\text{Residual SS(no dummies)} - \text{Residual SS}(D_p \text{ dummies used})]/p}{\text{Residual SS}(D_p \text{ dummies used})/(n-k-p)} \tag{8.7}$$

This will be distributed as $F_{p,n-k-p}$ under the null, see Gentleman and Wilk (1975). Let

$$e_p = D_p' e, \quad \text{then} \quad E(e_p) = 0 \quad \text{and} \quad \text{var}(e_p) = \sigma^2 D_p' \bar{P}_X D_p \tag{8.8}$$

Then one can show, (see problem 5), that

$$F = \frac{[e_p'(D_p'\bar{P}_X D_p)^{-1} e_p]/p}{[(n-k)s^2 - e_p'(D_p'\bar{P}_X D_p)^{-1} e_p]/(n-k-p)} \sim F_{p,n-k-p} \tag{8.9}$$

Another refinement comes from estimating the regression without the i-th observation:

$$\widehat{\beta}_{(i)} = [X_{(i)}' X_{(i)}]^{-1} X_{(i)}' y_{(i)} \tag{8.10}$$

where the (i) subscript notation indicates that the i-th observation has been deleted. Using the updating formula

$$(A - a'b)^{-1} = A^{-1} + A^{-1}a'(I - bA^{-1}a')^{-1}bA^{-1} \tag{8.11}$$

with $A = (X'X)$ and $a = b = x_i'$, one gets

$$[X_{(i)}' X_{(i)}]^{-1} = (X'X)^{-1} + (X'X)^{-1}x_i x_i'(X'X)^{-1}/(1 - h_{ii}) \tag{8.12}$$

Therefore

$$\widehat{\beta} - \widehat{\beta}_{(i)} = (X'X)^{-1}x_i e_i/(1 - h_{ii}) \tag{8.13}$$

Since the estimated coefficients are often of primary interest, (8.13) describes the change in the estimated regression coefficients that would occur if the i-th observation is deleted. Note that a high leverage observation with h_{ii} large will be influential in (8.13) only if the corresponding residual e_i is not small. Therefore, high leverage implies a potentially influential observation, but whether this potential is actually realized depends on y_i.

Alternatively, one can obtain this result from the augmented regression given in (8.5). Note that $P_{d_i} = d_i(d_i'd_i)^{-1}d_i' = d_i d_i'$ is an $n \times n$ matrix with 1 in the i-th diagonal position and 0 elsewhere. $\bar{P}_{d_i} = I_n - P_{d_i}$, has the effect when post-multiplied by a vector y of deleting the i-th observation. Hence, premultiplying (8.5) by $\bar{P}_{d_i}$ one gets

$$\bar{P}_{d_i} y = \begin{pmatrix} y_{(i)} \\ 0 \end{pmatrix} = \begin{pmatrix} X_{(i)} \\ 0 \end{pmatrix} \beta^* + \begin{pmatrix} u_{(i)} \\ 0 \end{pmatrix} \tag{8.14}$$

where the i-th observation is moved to the bottom of the data, without loss of generality. The last observation has no effect on the least squares estimate of β^* since both the dependent and independent variables are zero. This regression will yield $\widehat{\beta}^* = \widehat{\beta}_{(i)}$, and the i-th observation's residual is clearly zero. By the Frisch-Waugh-Lovell Theorem given in section 7.3, the least squares estimates and the residuals from (8.14) are numerically identical to those from (8.5).

Therefore, $\widehat{\beta}^* = \widehat{\beta}_{(i)}$ in (8.5) and the i-th observation residual from (8.5) must be zero. This implies that $\widehat{\varphi} = y_i - x_i'\widehat{\beta}_{(i)}$, and the fitted values from this regression are given by $\widehat{y} = X\widehat{\beta}_{(i)} + d_i\widehat{\varphi}$ whereas those from the original regression (7.1) are given by $X\widehat{\beta}$. The difference in residuals is therefore

$$e - e_{(i)} = X\widehat{\beta}_{(i)} + d_i\widehat{\varphi} - X\widehat{\beta} \tag{8.15}$$

premultiplying (8.15) by $\bar{P}_X$ and using the fact that $\bar{P}_X X = 0$, one gets $\bar{P}_X(e - e_{(i)}) = \bar{P}_X d_i\widehat{\varphi}$. But, $\bar{P}_X e = e$ and $\bar{P}_X e_{(i)} = e_{(i)}$, hence $\bar{P}_X d_i\widehat{\varphi} = e - e_{(i)}$. Premultiplying both sides by d_i' one gets $d_i'\bar{P}_X d_i\widehat{\varphi} = e_i$ since the i-th residual of $e_{(i)}$ from (8.5) is zero. By definition, $d_i'\bar{P}_X d_i = 1 - h_{ii}$, therefore

$$\widehat{\varphi} = e_i/(1 - h_{ii}) \tag{8.16}$$

premultiplying (8.15) by $(X'X)^{-1}X'$ one gets $0 = \widehat{\beta}_{(i)} - \widehat{\beta} + (X'X)^{-1}X'd_i\widehat{\varphi}$. This uses the fact that both residuals are orthogonal to X. Rearranging terms and substituting $\widehat{\varphi}$ from (8.16), one gets

$$\widehat{\beta} - \widehat{\beta}_{(i)} = (X'X)^{-1}x_i\widehat{\varphi} = (X'X)^{-1}x_i e_i/(1 - h_{ii})$$

as given in (8.13).

Note that $s^2_{(i)}$ given in (8.2) can now be written in terms of $\widehat{\beta}_{(i)}$:

$$s^2_{(i)} = \textstyle\sum_{t\neq i}(y_t - x_t'\widehat{\beta}_{(i)})^2/(n-k-1) \tag{8.17}$$

upon substituting (8.13) in (8.17) we get

$$\begin{aligned}(n-k-1)s^2_{(i)} &= \textstyle\sum_{t=1}^n \left(e_t + \dfrac{h_{it}e_i}{1-h_{ii}}\right)^2 - \dfrac{e_i^2}{(1-h_{ii})^2} \\ &= (n-k)s^2 + \dfrac{2e_i}{1-h_i}\textstyle\sum_{t=1}^n e_t h_{it} + \dfrac{e_i^2}{(1-h_{ii})^2}\sum_{t=1}^n h_{it}^2 - \dfrac{e_i^2}{(1-h_{ii})^2} \\ &= (n-k)s^2 - \dfrac{e_i^2}{1-h_{ii}} \end{aligned} \tag{8.18}$$

which is (8.2). This uses the fact that $He = 0$ and $H^2 = H$. Hence, $\sum_{t=1}^n e_t h_{it} = 0$ and $\sum_{t=1}^n h_{it}^2 = h_{ii}$.

To assess whether the change in $\widehat{\beta}_j$ (the j-th component of $\widehat{\beta}$) that results from the deletion of the i-th observation, is large or small, we scale by the variance of $\widehat{\beta}_j$, $\sigma^2(X'X)^{-1}_{jj}$. This is denoted by

$$DFBETAS_{ij} = (\widehat{\beta}_j - \widehat{\beta}_{j(i)})/s_{(i)}\sqrt{(X'X)^{-1}_{jj}} \tag{8.19}$$

Note that $s_{(i)}$ is used in order to make the denominator stochastically independent of the numerator in the Gaussian case. Absolute values of *DFBETAS* larger than 2 are considered influential. However, Belsley, Kuh, and Welsch (1980) suggest $2/\sqrt{n}$ as a size-adjusted cutoff. In fact, it would be most unusual for the removal of a single observation from a sample of 100

or more to result in a change in any estimate by two or more standard errors. The size-adjusted cutoff tend to expose approximately the same proportion of potentially influential observations, regardless of sample size. The size-adjusted cutoff is particularly important for large data sets.

In case of Normality, it can also be useful to look at the change in the t-statistics, as a means of assessing the sensitivity of the regression output to the deletion of the i-th observation:

$$DFSTAT_{ij} = \frac{\widehat{\beta}_j}{s\sqrt{(X'X)^{-1}_{jj}}} - \frac{\widehat{\beta}_{j(i)}}{s_{(i)}\sqrt{(X'_{(i)}X_{(i)})^{-1}_{jj}}} \tag{8.20}$$

Another way to summarize coefficient changes and gain insight into forecasting effects when the i-th observation is deleted is to look at the change in fit, defined as

$$DFFIT_i = \widehat{y}_i - \widehat{y}_{(i)} = x'_i[\widehat{\beta} - \widehat{\beta}_{(i)}] = h_{ii}e_i/(1-h_{ii}) \tag{8.21}$$

where the last equality is obtained from (8.13).

We scale this measure by the variance of $\widehat{y}_{(i)}$, i.e., $\sigma\sqrt{h_{ii}}$, giving

$$DFFITS_i = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}} = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} e_i^* \tag{8.22}$$

where σ has been estimated by $s_{(i)}$ and e_i^* denotes the externally studentized residual given in (8.3). Values of *DFFITS* larger than 2 in absolute value are considered influential. A size-adjusted cutoff for *DFFITS* suggested by Belsley, Kuh and Welsch (1980) is $2\sqrt{k/n}$.

In (8.3), the studentized residual e_i^* was interpreted as a t-statistic that tests for the significance of the coefficient φ of d_i, the dummy variable which takes the value 1 for the i-th observation and 0 otherwise, in the regression of y on X and d_i. This can now be easily proved as follows:

Consider the Chow test for the significance of φ. The $RRSS = (n-k)s^2$, the $URSS = (n-k-1)s^2_{(i)}$ and the Chow F-test described in (4.17) becomes

$$F_{1,n-k-1} = \frac{[(n-k)s^2 - (n-k-1)s^2_{(i)}]/1}{(n-k-1)s^2_{(i)}/(n-k-1)} = \frac{e_i^2}{s^2_{(i)}(1-h_{ii})} \tag{8.23}$$

The square root of (8.23) is $e_i^* \sim t_{n-k-1}$. These studentized residuals provide a better way to examine the information in the residuals, but they do not tell the whole story, since some of the most influential data points can have small e_i^* (and very small e_i).

One overall measure of the impact of the i-th observation on the estimated regression coefficients is Cook's (1977) distance measure D_i^2. Recall, that the confidence region for all k regression coefficients is $(\widehat{\beta}-\beta)'X'X(\widehat{\beta}-\beta)/ks^2 \sim F(k, n-k)$. Cook's (1977) distance measure D_i^2 uses the same structure for measuring the combined impact of the differences in the estimated regression coefficients when the i-th observation is deleted:

$$D_i^2(s) = (\widehat{\beta} - \widehat{\beta}_{(i)})'X'X(\widehat{\beta} - \widehat{\beta}_{(i)})/ks^2 \tag{8.24}$$

Even though $D_i^2(s)$ does not follow the above F-distribution, Cook suggests computing the percentile value from this F-distribution and declaring an influential observation if this percentile value $\geq 50\%$. In this case, the distance between $\widehat{\beta}$ and $\widehat{\beta}_{(i)}$ will be large, implying that the i-th

observation has a substantial influence on the fit of the regression. Cook's distance measure can be equivalently computed as:

$$D_i^2(s) = \frac{e_i^2}{ks^2}\left(\frac{h_{ii}}{(1-h_{ii})^2}\right) \tag{8.25}$$

$D_i^2(s)$ depends on e_i and h_{ii}; the larger e_i or h_{ii} the larger is $D_i^2(s)$. Note the relationship between Cook's $D_i^2(s)$ and Belsley, Kuh, and Welsch (1980) $DFFITS_i(\sigma)$ in (8.22), i.e.,

$$DFFITS_i(\sigma) = \sqrt{k}D_i(\sigma) = (\widehat{y}_i - x_i'\widehat{\beta}_{(i)})/(\sigma\sqrt{h_{ii}})$$

Belsley, Kuh, and Welsch (1980) suggest nominating *DFFITS* based on $s_{(i)}$ exceeding $2\sqrt{k/n}$ for special attention. Cook's 50 percentile recommendation is equivalent to $DFFITS > \sqrt{k}$, which is more conservative, see Velleman and Welsch (1981).

Next, we study the influence of the i-th observation deletion on the covariance matrix of the regression coefficients. One can compare the two covariance matrices using the ratio of their determinants:

$$COVRATIO_i = \frac{\det(s_{(i)}^2[X_{(i)}'X_{(i)}]^{-1})}{\det(s^2[X'X]^{-1})} = \frac{s_{(i)}^{2k}}{s^{2k}}\left(\frac{\det[X_{(i)}'X_{(i)}]^{-1}}{\det[X'X]^{-1}}\right) \tag{8.26}$$

Using the fact that

$$\det[X_{(i)}'X_{(i)}] = (1-h_{ii})\det[X'X] \tag{8.27}$$

see problem 8, one obtains

$$COVRATIO_i = \left(\frac{s_{(i)}^2}{s^2}\right)^k \times \frac{1}{1-h_{ii}} = \frac{1}{\left(\frac{n-k-1}{n-k} + \frac{e_i^{*2}}{n-k}\right)^k (1-h_{ii})} \tag{8.28}$$

where the last equality follows from (8.18) and the definition of e_i^* in (8.3). Values of *COVRATIO* not near unity identify possible influential observations and warrant further investigation. Belsley, Kuh and Welsch (1980) suggest investigating points with $|COVRATIO - 1|$ near to or larger than $3k/n$. The *COVRATIO* depends upon both h_{ii} and e_i^{*2}. In fact, from (8.28), *COVRATIO* is large when h_{ii} is large and small when e_i^* is large. The two factors can offset each other, that is why it is important to look at h_{ii} and e_i^* separately as well as in combination as in *COVRATIO*.

Finally, one can look at how the variance of $\widehat{y}_i$ changes when an observation is deleted.

$$\text{var}(\widehat{y}_i) = s^2 h_{ii} \qquad \text{and} \qquad \text{var}(\widehat{y}_{(i)}) = \text{var}(x_i'\widehat{\beta}_{(i)}) = s_{(i)}^2(h_{ii}/(1-h_{ii}))$$

and the ratio is

$$FVARATIO_i = s_{(i)}^2/s^2(1-h_{ii}) \tag{8.29}$$

This expression is similar to *COVRATIO* except that $[s_{(i)}^2/s^2]$ is not raised to the k-th power. As a diagnostic measure it will exhibit the same patterns of behavior with respect to different configurations of h_{ii} and the studentized residual as described for *COVRATIO*.

Table 8.1 Cigarette Regression

Dependent Variable: LNC
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	0.50098	0.25049	9.378	0.0004
Error	43	1.14854	0.02671		
C Total	45	1.64953			

Root MSE	0.16343	R-square	0.3037
Dep Mean	4.84784	Adj R-sq	0.2713
C.V.	3.37125		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	4.299662	0.90892571	4.730	0.0001
LNP	1	–1.338335	0.32460147	–4.123	0.0002
LNY	1	0.172386	0.19675440	0.876	0.3858

Example 1: For the cigarette data given in Table 3.2, Table 8.1 gives the SAS least squares regression for log$C$ on log$P$ and log$Y$.

$$\begin{array}{rccccc} \log C = & 4.30 & - & 1.34 \ \log P + & 0.172 \ \log Y + \text{residuals} \\ & (0.909) & & (0.325) & & (0.197) \end{array}$$

The standard error of the regression is $s = 0.16343$ and $\bar{R}^2 = 0.271$. Table 8.2 gives the data along with the predicted values of log$C$, the least squares residuals e, the internal studentized residuals $\widetilde{e}$ given in (8.1), the externally studentized residuals e^* given in (8.3), the Cook statistic given in (8.25), the leverage of each observation h, the *DFFITS* given in (8.22) and the *COVRATIO* given in (8.28).

Using the leverage column, one can identify four potential observations with high leverage, i.e., greater than $2\bar{h} = 2k/n = 6/46 = 0.13043$. These are the observations belonging to the following states: Connecticut (CT), Kentucky (KY), New Hampshire (NH) and New Jersey (NJ) with leverage $0.13535, 0.19775, 0.13081$ and 0.13945, respectively. Note that the corresponding OLS residuals are $-0.078, 0.234, 0.160$ and -0.059, which are not necessarily large. The internally studentized residuals are computed using equation (8.1). For KY this gives

$$\widetilde{e}_{KY} = \frac{e_{KY}}{s\sqrt{1-h_{KY}}} = \frac{0.23428}{0.16343\sqrt{1-0.19775}} = 1.6005$$

From Table 8.2, two observations with a high internally studentized residuals are those belonging to Arkansas (AR) and Utah (UT) with values of 2.102 and -2.679 respectively, both larger than 2 in absolute value.

The externally studentized residuals are computed from (8.3). For KY, we first compute $s^2_{(KY)}$, the MSE from the regression computed without the KY observation. From (8.2), this is

given by

$$\begin{aligned} s^2_{(KY)} &= \frac{(n-k)s^2 - e^2_{KY}/(1-h_{KY})}{(n-k-1)} \\ &= \frac{(46-3)(0.16343)^2 - (0.23428)^2/(1-0.19775)}{(46-3-1)} = 0.025716 \end{aligned}$$

From (8.3) we get

$$e^*_{(KY)} = \frac{e_{KY}}{s_{(KY)}\sqrt{1-h_{KY}}} = \frac{0.23428}{0.16036\sqrt{1-0.19775}} = 1.6311$$

This externally studentized residual is distributed as a t-statistic with 42 degrees of freedom. However, e^*_{KY} does not exceed 2 in absolute value. Again, e^*_{AR} and e^*_{UT} are 2.193 and -2.901 both larger than 2 in absolute value. From (8.13), the change in the regression coefficients due to the omission of the KY observation is given by

$$\widehat{\beta} - \widehat{\beta}_{(KY)} = (X'X)^{-1} x_{KY} e_{KY}/(1-h_{KY})$$

Using the fact that

$$(X'X)^{-1} = \begin{bmatrix} 30.929816904 & 4.8110214655 & -6.679318415 \\ 4.81102114655 & 3.9447686638 & -1.177208398 \\ -6.679318415 & -1.177208398 & 1.4493372835 \end{bmatrix}$$

and $x'_{KY} = (1, -0.03260, 4.64937)$ with $e_{KY} = 0.23428$ and $h_{KY} = 0.19775$ one gets

$$(\widehat{\beta} - \widehat{\beta}_{(KY)})' = (-0.082249, -0.230954, 0.028492)$$

In order to assess whether this change is large or small, we compute DFBETAS given in (8.19). For the KY observation, these are given by

$$DFBETAS_{KY,1} = \frac{\widehat{\beta}_1 - \widehat{\beta}_{1(KY)}}{s_{(KY)}\sqrt{(X'X)^{-1}_{11}}} = \frac{-0.082449}{0.16036\sqrt{30.9298169}} = -0.09222$$

Similarly, $DFBETAS_{KY,2} = -0.7251$ and $DFBETAS_{KY,3} = 0.14758$. These are not larger than 2 in absolute value. However, $DFBETAS_{KY,2}$ is larger than $2/\sqrt{n} = 2/\sqrt{46} = 0.2949$ in absolute value. This is the size-adjusted cutoff recommended by Belsley, Kuh and Welsch (1980) for large n.

The change in the fit due to the omission of the KY observation is given by (8.21). In fact,

$$\begin{aligned} DFFIT_{KY} &= \widehat{y}_{KY} - \widehat{y}_{(KY)} = x'_{KY}[\widehat{\beta} - \widehat{\beta}_{(KY)}] \\ &= (1, -0.03260, 4.64937) \begin{pmatrix} -0.082249 \\ -0.230954 \\ -0.028492 \end{pmatrix} = 0.05775 \end{aligned}$$

or simply

$$DFFIT_{KY} = \frac{h_{KY} e_{KY}}{(1-h_{KY})} = \frac{(0.19775)(0.23428)}{1-0.19775} = 0.05775$$

Scaling it by the variance of $\widehat{y}_{(KY)}$ we get from (8.22)

$$DFFITS_{KY} = \left(\frac{h_{KY}}{1-h_{KY}}\right)^{1/2} e^*_{KY} = \left(\frac{0.19775}{1-0.19775}\right)^{1/2} (1.6311) = 0.8098$$

This is not larger than 2 in absolute value, but it is larger than the size-adjusted cutoff of $2\sqrt{k/n} = 2\sqrt{3/46} = 0.511$. Note also that both $DFFITS_{AR} = 0.667$ and $DFFITS_{UT} = -0.888$ are larger than 0.511 in absolute value.

Cook's distance measure is given in (8.25) and for KY can be computed as

$$D^2_{KY}(s) = \frac{e^2_{KY}}{ks^2}\left(\frac{h_{KY}}{(1-h_{KY})^2}\right) = \left(\frac{(0.23428)^2}{3(0.16343)^2}\right)\left(\frac{0.19775}{(1-0.19775)^2}\right) = 0.21046$$

The other two large Cook's distance measures are $DR^2_{AR}(s) = 0.13623$ and $D^2_{UT}(s) = 0.22399$, respectively. *COVRATIO* omitting the KY observation can be computed from (8.28) as

$$COVRATIO_{KY} = \left(\frac{s^2_{(KY)}}{s^2}\right)^k \frac{1}{1-h_{KY}} = \left(\frac{0.025716}{(0.16343)^2}\right)^3 \left(\frac{1}{(1-0.019775)}\right) = 1.1125$$

which means that $COVRATIO_{KY} - 1/ = 0.1125$ is less than $3k/n = 9/46 = 0.1956$.

Finally, *FVARATIO* omitting the KY observation can be computed from (8.29) as

$$FVARATIO_{KY} = \frac{s^2_{(KY)}}{s^2(1-h_{KY})} = \frac{0.025716}{(0.16343)^2(1-0.19775)} = 1.2001$$

By several diagnostic measures, AR, KY and UT are influential observations that deserve special attention. The first two states are characterized with large sales of cigarettes. KY is a producer state with a very low price on cigarettes, while UT is a low consumption state due to its high percentage of Mormon population (a religion that forbids smoking). Table 8.3 gives the predicted consumption along with the 95% confidence band, the OLS residuals, and the internalized student residuals, Cook's D-statistic and a plot of these residuals. This last plot highlights the fact that AR, UT and KY have large studentized residuals.

8.2 Recursive Residuals

In Section 8.1, we showed that the least squares residuals are heteroskedastic with non-zero covariances, even when the true disturbances have a scalar covariance matrix. This section studies recursive residuals which are a set of linear unbiased residuals with a scalar covariance matrix. They are independent and identically distributed when the true disturbances themselves are independent and identically distributed.[2] These residuals are natural in time-series regressions and can be constructed as follows:

1. Choose the first $t \geq k$ observations and compute $\widehat{\beta}_t = (X_t'X_t)^{-1}X_t'Y_t$ where X_t denotes the $t \times k$ matrix of t observations on k variables and $Y_t' = (y_1, \ldots, y_t)$. The recursive residuals are basically standardized one-step ahead forecast residuals:

$$w_{t+1} = (y_{t+1} - x_{t+1}'\widehat{\beta}_t)/\sqrt{1 + x_{t+1}'(X_t'X_t)^{-1}x_{t+1}} \tag{8.30}$$

Table 8.2 Diagnostic Statistics for the Cigarettes Example

OBS	STATE	LNC	LNP	LNY	PREDICTED	e	$\tilde{e}$	$e^{\star}$	Cook's D	Leverage	DFFITS	COVRATIO
1	AL	4.96213	0.20487	4.64039	4.8254	0.1367	0.857	0.8546	0.012	0.0480	0.1919	1.0704
2	AZ	4.66312	0.16640	4.68389	4.8844	–0.2213	–1.376	–1.3906	0.021	0.0315	–0.2508	0.9681
3	AR	5.10709	0.23406	4.59435	4.7784	0.3287	2.102	2.1932	0.136	0.0847	0.6670	0.8469
4	CA	4.50449	0.36399	4.88147	4.6540	–0.1495	–0.963	–0.9623	0.033	0.0975	–0.3164	1.1138
5	CT	4.66983	0.32149	5.09472	4.7477	–0.0778	–0.512	–0.5077	0.014	0.1354	–0.2009	1.2186
6	DE	5.04705	0.21929	4.87087	4.8458	0.2012	1.252	1.2602	0.018	0.0326	0.2313	0.9924
7	DC	4.65637	0.28946	5.05960	4.7845	–0.1281	–0.831	–0.8280	0.029	0.1104	–0.2917	1.1491
8	FL	4.80081	0.28733	4.81155	4.7446	0.0562	0.352	0.3482	0.002	0.0431	0.0739	1.1118
9	GA	4.97974	0.12826	4.73299	4.9439	0.0358	0.224	0.2213	0.001	0.0402	0.0453	1.1142
10	ID	4.74902	0.17541	4.64307	4.8653	–0.1163	–0.727	–0.7226	0.008	0.0413	–0.1500	1.0787
11	IL	4.81445	0.24806	4.90387	4.8130	0.0014	0.009	0.0087	0.000	0.0399	0.0018	1.1178
12	IN	5.11129	0.08992	4.72916	4.9946	0.1167	0.739	0.7347	0.013	0.0650	0.1936	1.1046
13	IA	4.80857	0.24081	4.74211	4.7949	0.0137	0.085	0.0843	0.000	0.0310	0.0151	1.1070
14	KS	4.79263	0.21642	4.79613	4.8368	–0.0442	–0.273	–0.2704	0.001	0.0223	–0.0408	1.0919
15	KY	5.37906	–0.03260	4.64937	5.1448	0.2343	1.600	1.6311	0.210	0.1977	0.8098	1.1126
16	LA	4.98602	0.23856	4.61461	4.7759	0.2101	1.338	1.3504	0.049	0.0761	0.3875	1.0224
17	ME	4.98722	0.29106	4.75501	4.7298	0.2574	1.620	1.6527	0.051	0.0553	0.4000	0.9403
18	MD	4.77751	0.12575	4.94692	4.9841	–0.2066	–1.349	–1.3624	0.084	0.1216	–0.5070	1.0731
19	MA	4.73877	0.22613	4.99998	4.8590	–0.1202	–0.769	–0.7653	0.018	0.0856	–0.2341	1.1258
20	MI	4.94744	0.23067	4.80620	4.8195	0.1280	0.792	0.7890	0.005	0.0238	0.1232	1.0518
21	MN	4.69589	0.34297	4.81207	4.6702	0.0257	0.165	0.1627	0.001	0.0864	0.0500	1.1724
22	MS	4.93990	0.13638	4.52938	4.8979	0.0420	0.269	0.2660	0.002	0.0883	0.0828	1.1712
23	MO	5.06430	0.08731	4.78189	5.0071	0.0572	0.364	0.3607	0.004	0.0787	0.1054	1.1541
24	MT	4.73313	0.15303	4.70417	4.9058	–0.1727	–1.073	–1.0753	0.012	0.0312	–0.1928	1.0210
25	NE	4.77558	0.18907	4.79671	4.8735	–0.0979	–0.607	–0.6021	0.003	0.0243	–0.0950	1.0719
26	NV	4.96642	0.32304	4.83816	4.7014	0.2651	1.677	1.7143	0.065	0.0646	0.4504	0.9366
27	NH	5.10990	0.15852	5.00319	4.9500	0.1599	1.050	1.0508	0.055	0.1308	0.4076	1.1422
28	NJ	4.70633	0.30901	5.10268	4.7657	–0.0594	–0.392	–0.3879	0.008	0.1394	–0.1562	1.2337
29	NM	4.58107	0.16458	4.58202	4.8693	–0.2882	–1.823	–1.8752	0.076	0.0639	–0.4901	0.9007
30	NY	4.66496	0.34701	4.96075	4.6904	–0.0254	–0.163	–0.1613	0.001	0.0888	–0.0503	1.1755
31	ND	4.58237	0.18197	4.69163	4.8649	–0.2825	–1.755	–1.7999	0.031	0.0295	–0.3136	0.8848
32	OH	4.97952	0.12889	4.75875	4.9475	0.0320	0.200	0.1979	0.001	0.0423	0.0416	1.1174
33	OK	4.72720	0.19554	4.62730	4.8356	–0.1084	–0.681	–0.6766	0.008	0.0505	–0.1560	1.0940
34	PA	4.80363	0.22784	4.83516	4.8282	–0.0246	–0.153	–0.1509	0.000	0.0257	–0.0245	1.0997
35	RI	4.84693	0.30324	4.84670	4.7293	0.1176	0.738	0.7344	0.010	0.0504	0.1692	1.0876
36	SC	5.07801	0.07944	4.62549	4.9907	0.0873	0.555	0.5501	0.008	0.0725	0.1538	1.1324
37	SD	4.81545	0.13139	4.67747	4.9301	–0.1147	–0.716	–0.7122	0.007	0.0402	–0.1458	1.0786
38	TN	5.04939	0.15547	4.72525	4.9062	0.1432	0.890	0.8874	0.008	0.0294	0.1543	1.0457
39	TX	4.65398	0.28196	4.73437	4.7384	–0.0845	–0.532	–0.5271	0.005	0.0546	–0.1267	1.1129
40	UT	4.40859	0.19260	4.55586	4.8273	–0.4187	–2.679	–2.9008	0.224	0.0856	–0.8876	0.6786
41	VT	5.08799	0.18018	4.77578	4.8818	0.2062	1.277	1.2869	0.014	0.0243	0.2031	0.9794
42	VA	4.93065	0.11818	4.85490	4.9784	–0.0478	–0.304	–0.3010	0.003	0.0773	–0.0871	1.1556
43	WA	4.66134	0.35053	4.85645	4.6677	–0.0064	–0.041	–0.0404	0.000	0.0866	–0.0124	1.1747
44	WV	4.82454	0.12008	4.56859	4.9265	–0.1020	–0.647	–0.6429	0.011	0.0709	–0.1777	1.1216
45	WI	4.83026	0.22954	4.75826	4.8127	0.0175	0.109	0.1075	0.000	0.0254	0.0174	1.1002
46	WY	5.00087	0.10029	4.71169	4.9777	0.0232	0.146	0.1444	0.000	0.0555	0.0350	1.1345

Table 8.3 Regression of Real Per-Capita Consumption of Cigarettes

Dep Obs	Var LNC	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict	Upper95% Predict	Std Err Residual	Student Residual	Residual	−2 −1 0 1 2	Cook's D
1	4.9621	4.8254	0.036	4.7532	4.8976	4.4880	5.1628	0.1367	0.159	0.857	*	0.012
2	4.6631	4.8844	0.029	4.8259	4.9429	4.5497	5.2191	–0.2213	0.161	–1.376	**	0.021
3	5.1071	4.7784	0.048	4.6825	4.8743	4.4351	5.1217	0.3287	0.156	2.102	****	0.136
4	4.5045	4.6540	0.051	4.5511	4.7570	4.3087	4.9993	–0.1495	0.155	–0.963	*	0.033
5	4.6698	4.7477	0.060	4.6264	4.8689	4.3965	5.0989	–0.0778	0.152	–0.512	*	0.014
6	5.0471	4.8458	0.030	4.7863	4.9053	4.5109	5.1808	0.2012	0.161	1.252	**	0.018
7	4.6564	4.7845	0.054	4.6750	4.8940	4.4372	5.1318	–0.1281	0.154	–0.831	*	0.029
8	4.8008	4.7446	0.034	4.6761	4.8130	4.4079	5.0812	0.0562	0.160	0.352		0.002
9	4.9797	4.9439	0.033	4.8778	5.0100	4.6078	5.2801	0.0358	0.160	0.224		0.001
10	4.7490	4.8653	0.033	4.7983	4.9323	4.5290	5.2016	–0.1163	0.160	–0.727	*	0.008
11	4.8145	4.8130	0.033	4.7472	4.8789	4.4769	5.1491	0.00142	0.160	0.009		0.000
12	5.1113	4.9946	0.042	4.9106	5.0786	4.6544	5.3347	0.1167	0.158	0.739	*	0.013
13	4.8086	4.7949	0.029	4.7368	4.8529	4.4602	5.1295	0.0137	0.161	0.085		0.000
14	4.7926	4.8368	0.024	4.7876	4.8860	4.5036	5.1701	–0.0442	0.162	–0.273		0.001
15	5.3791	5.1448	0.073	4.9982	5.2913	4.7841	5.5055	0.2343	0.146	1.600	***	0.210
16	4.9860	4.7759	0.045	4.6850	4.8668	4.4340	5.1178	0.2101	0.157	1.338	**	0.049
17	4.9872	4.7298	0.038	4.6523	4.8074	4.3912	5.0684	0.2574	0.159	1.620	***	0.051
18	4.7775	4.9841	0.057	4.8692	5.0991	4.6351	5.3332	–0.2066	0.153	–1.349	**	0.084
19	4.7388	4.8590	0.048	4.7625	4.9554	4.5155	5.2024	–0.1202	0.156	–0.769	*	0.018
20	4.9474	4.8195	0.025	4.7686	4.8703	4.4860	5.1530	0.1280	0.161	0.792	*	0.005
21	4.6959	4.6702	0.048	4.5733	4.7671	4.3267	5.0137	0.0257	0.156	0.165		0.001
22	4.9399	4.8979	0.049	4.8000	4.9959	4.5541	5.2418	0.0420	0.156	0.269		0.002
23	5.0643	5.0071	0.046	4.9147	5.0996	4.6648	5.3495	0.0572	0.157	0.364		0.004
24	4.7331	4.9058	0.029	4.8476	4.9640	4.5711	5.2405	–0.1727	0.161	–1.073	**	0.012
25	4.7756	4.8735	0.025	4.8221	4.9249	4.5399	5.2071	–0.0979	0.161	–0.607	*	0.003
26	4.9664	4.7014	0.042	4.6176	4.7851	4.3613	5.0414	0.2651	0.158	1.677	***	0.065
27	5.1099	4.9500	0.059	4.8308	5.0692	4.5995	5.3005	0.1599	0.152	1.050	**	0.055
28	4.7063	4.7657	0.061	4.6427	4.8888	4.4139	5.1176	–0.0594	0.152	–0.392		0.008
29	4.5811	4.8693	0.041	4.7859	4.9526	4.5293	5.2092	–0.2882	0.158	–1.823	***	0.076
30	4.6650	4.6904	0.049	4.5922	4.7886	4.3465	5.0343	–0.0254	0.156	–0.163		0.001
31	4.5824	4.8649	0.028	4.8083	4.9215	4.5305	5.1993	–0.2825	0.161	–1.755	***	0.031
32	4.9795	4.9475	0.034	4.8797	5.0153	4.6110	5.2840	0.0320	0.160	0.200		0.001
33	4.7272	4.8356	0.037	4.7616	4.9097	4.4978	5.1735	–0.1084	0.159	–0.681	*	0.008
34	4.8036	4.8282	0.026	4.7754	4.8811	4.4944	5.1621	–0.0246	0.161	–0.153		0.000
35	4.8469	4.7293	0.037	4.6553	4.8033	4.3915	5.0671	0.1176	0.159	0.738	*	0.010
36	5.0780	4.9907	0.044	4.9020	5.0795	4.6494	5.3320	0.0873	0.157	0.555	*	0.008
37	4.8155	4.9301	0.033	4.8640	4.9963	4.5940	5.2663	–0.1147	0.160	–0.716	*	0.007
38	5.0494	4.9062	0.028	4.8497	4.9626	4.5718	5.2406	0.1432	0.161	0.890	*	0.008
39	4.6540	4.7384	0.038	4.6614	4.8155	4.4000	5.0769	–0.0845	0.159	–0.532	*	0.005
40	4.4086	4.8273	0.048	4.7308	4.9237	4.4839	5.1707	–0.4187	0.156	–2.679	*****	0.224
41	5.0880	4.8818	0.025	4.8304	4.9332	4.5482	5.2154	0.2062	0.161	1.277	**	0.014
42	4.9307	4.9784	0.045	4.8868	5.0701	4.6363	5.3205	–0.0478	0.157	–0.304		0.003
43	4.6613	4.6677	0.048	4.5708	4.7647	4.3242	5.0113	–0.00638	0.156	–0.041		0.000
44	4.8245	4.9265	0.044	4.8387	5.0143	4.5854	5.2676	–0.1020	0.158	–0.647	*	0.011
45	4.8303	4.8127	0.026	4.7602	4.8653	4.4790	5.1465	0.0175	0.161	0.109		0.000
46	5.0009	4.9777	0.039	4.9000	5.0553	4.6391	5.3163	0.0232	0.159	0.146		0.000

Sum of Residuals	0
Sum of Squared Residuals	1.1485
Predicted Resid SS (Press)	1.3406

2. Add the $(t+1)$-th observation to the data and obtain $\widehat{\beta}_{t+1} = (X'_{t+1}X_{t+1})^{-1}X'_{t+1}Y_{t+1}$. Compute w_{t+2}.

3. Repeat step 2, adding one observation at a time. In time-series regressions, one usually starts with the first k-observations and obtain $(T-k)$ forward recursive residuals. These recursive residuals can be computed using the updating formula given in (8.11) with $A = (X'_tX_t)$ and $a = -b = x'_{t+1}$. Therefore,

$$(X'_{t+1}X_{t+1})^{-1} = (X'_tX_t)^{-1} - (X'_tX_t)^{-1}x_{t+1}x'_{t+1}(X'_tX_t)^{-1}/[1+x'_{t+1}(X'_tX_t)^{-1}x_{t+1}] \quad (8.31)$$

and only $(X'_tX_t)^{-1}$ have to be computed. Also,

$$\widehat{\beta}_{t+1} = \widehat{\beta}_t + (X'_tX_t)^{-1}x_{t+1}(y_{t+1} - x'_{t+1}\widehat{\beta}_t)/f_{t+1} \quad (8.32)$$

where $f_{t+1} = 1 + x'_{t+1}(X'_tX_t)^{-1}x_{t+1}$, see problem 13.

Alternatively, one can compute these residuals by regressing Y_{t+1} on X_{t+1} and d_{t+1} where $d_{t+1} = 1$ for the $(t+1)$-th observation, and zero otherwise, see equation (8.5). The estimated coefficient of d_{t+1} is the numerator of w_{t+1}. The standard error of this estimate is s_{t+1} times the denominator of w_{t+1}, where s_{t+1} is the standard error of this regression. Hence, w_{t+1} can be retrieved as s_{t+1} multiplied by the t-statistic corresponding to d_{t+1}. This computation has to be performed sequentially, in each case generating the corresponding recursive residual. This may be computationally inefficient, but it is simple to generate using regression packages.

It is obvious from (8.30) that if $u_t \sim \text{IIN}(0, \sigma^2)$, then w_{t+1} has zero mean and $\text{var}(w_{t+1}) = \sigma^2$. Furthermore, w_{t+1} is linear in the y's. Therefore, it is normally distributed. It remains to show that the recursive residuals are independent. Given normality, it is sufficient to show that

$$\text{cov}(w_{t+1}, w_{s+1}) = 0 \qquad \text{for} \qquad t \neq s;\ t, s = k, \ldots, T-1 \quad (8.33)$$

This is left as an exercise for the reader, see problem 13.

Alternatively, one can express the $T-k$ vector of recursive residuals as $w = Cy$ where C is of dimension $(T-k) \times T$ as follows:

$$C = \begin{bmatrix} -\frac{x'_{k+1}(X'_kX_k)^{-1}X'_k}{\sqrt{f_{k+1}}} & \frac{1}{\sqrt{f_{k+1}}} & & 0....0 \\ \vdots & & \ddots & \\ -\frac{x'_t(X'_{t-1}X_{t-1})^{-1}X'_{t-1}}{\sqrt{f_t}} & & \frac{1}{\sqrt{f_t}} & 0....0 \\ \vdots & & & \ddots \\ -\frac{x'_T(X'_{T-1}X_{T-1})^{-1}X'_{T-1}}{\sqrt{f_T}} & & & \frac{1}{\sqrt{f_T}} \end{bmatrix} \quad (8.34)$$

Problem 14 asks the reader to verify that $w = Cy$, using (8.30). Also, that the matrix C satisfies the following properties:

$$(i)\ CX = 0 \qquad (ii)\ CC' = I_{T-k} \qquad (iii)\ C'C = \bar{P}_X \quad (8.35)$$

This means that the recursive residuals w are (LUS) linear in y, unbiased with mean zero and have a scalar variance-covariance matrix: $\text{var}(w) = CE(uu')C' = \sigma^2 I_{T-k}$. Property (iii) also

means that $w'w = y'C'Cy = y'\bar{P}_X y = e'e$. This means that the sum of squares of $(T-k)$ recursive residuals is equal to the sum of squares of T least squares residuals. One can also show from (8.32) that

$$RSS_{t+1} = RSS_t + w_{t+1}^2 \qquad \text{for} \qquad t = k, \ldots, T-1 \tag{8.36}$$

where $RSS_t = (Y_t - X_t\widehat{\beta}_t)'(Y_t - X_t\widehat{\beta}_t)$, see problem 14. Note that for $t = k$; $RSS = 0$, since with k observations one gets a perfect fit and zero residuals. Therefore

$$RSS_T = \sum_{t=k+1}^{T} w_t^2 = \sum_{t=1}^{T} e_t^2 \tag{8.37}$$

Applications of Recursive Residuals

Recursive residuals have been used in several important applications:

(1) Harvey (1976) used these recursive residuals to give an alternative proof of the fact that Chow's post-sample *predictive test* has an F-distribution. Recall, from Chapter 7, that when the second sample n_2 had fewer than k observations, Chow's test becomes

$$F = \frac{(e'e - e_1'e_1)/n_2}{e_1'e_1/(n_1-k)} \sim F(n_2, n_1-k) \tag{8.38}$$

where $e'e = RSS$ from the total sample ($n_1 + n_2 = T$ observations), and $e_1'e_1 = RSS$ from the first n_1 observations. Recursive residuals can be computed for $t = k+1, \ldots, n_1$, and continued on for the extra n_2 observations. From (8.36) we have

$$e'e = \sum_{t=k+1}^{n_1+n_2} w_t^2 \qquad \text{and} \qquad e_1'e_1 = \sum_{t=k+1}^{n_1} w_t^2 \tag{8.39}$$

Therefore,

$$F = \frac{\sum_{t=n_1+1}^{n_1+n_2} w_t^2/n_2}{\sum_{t=k+1}^{n_1} w_t^2/(n_1-k)} \tag{8.40}$$

But the w_t's are $\sim$IIN$(0, \sigma^2)$ under the null, therefore the F-statistic in (8.38) is a ratio of two independent chi-squared variables, each divided by the appropriate degrees of freedom. Hence, $F \sim F(n_2, n_1-k)$ under the null, see Chapter 2.

(2) Harvey and Phillips (1974) used recursive residuals to test the null hypothesis of homoskedasticity. If the alternative hypothesis is that σ_i^2 varies with X_j, the proposed test is as follows:

1) Order the data according to X_j and choose a base of at least k observations from among the central observations.

2) From the first m observations compute the vector of recursive residuals w_1 using the base constructed in step 1. Also, compute the vector of recursive residuals w_2 from the last m observations. The maximum m can be is $(T-k)/2$.

3) Under the null hypothesis, it follows that

$$F = w_2'w_2/w_1'w_1 \sim F_{m,m} \tag{8.41}$$

Harvey and Phillips suggest setting m at approximately $(n/3)$ provided $n > 3k$. This test has the advantage over the Goldfeld-Quandt test in that if one wanted to test whether σ_i^2 varies with some other variable X_s, one could simply regroup the existing recursive residuals according to low and high values of X_s and compute (8.41) afresh, whereas the Goldfeld-Quandt test would require the computation of two new regressions.

(3) Phillips and Harvey (1974) suggest using the recursive residuals to test the null hypothesis of no serial correlation using a modified von Neuman ratio:

$$MVNR = \frac{\sum_{t=k+2}^{T}(w_t - w_{t-1})^2/(T-k-1)}{\sum_{t=k+1}^{T} w_t^2/(T-k)} \tag{8.42}$$

This is the ratio of the mean-square successive difference to the variance. It is arithmetically closely related to the DW statistic, but given that $w \sim N(0, \sigma^2 I_{T-k})$ one has an exact test available and no inconclusive regions. Phillips and Harvey (1974) provide tabulations of the significance points. If the sample size is large, a satisfactory approximation is obtained from a normal distribution with mean 2 and variance $4/(T-k)$.

(4) Harvey and Collier (1977) suggest a test for functional misspecification based on recursive residuals. This is based on the fact that $w \sim N(0, \sigma^2 I_{T-k})$. Therefore,

$$\bar{w}/(s_w/\sqrt{T-k}) \sim t_{T-k-1} \tag{8.43}$$

where $\bar{w} = \sum_{t=k+1}^{T} w_t/(T-k)$ and $s_w^2 = \sum_{t=k+1}^{T}(w_t - \bar{w})^2/(T-k-1)$. Suppose that the true functional form relating y to a single explanatory variable X is concave (convex) and the data are ordered by X. A simple linear regression is estimated by regressing y on X. The recursive residuals would be expected to be mainly negative (positive) and the computed t-statistic will be large in absolute value. When there are multiple X's, one could carry out this test based on any single explanatory variable. Since several specification errors might have a self-cancelling effect on the recursive residuals, this test is not likely to be very effective in multivariate situations. Recently, Wu (1993) suggested performing this test using the following augmented regression:

$$y = X\beta + z\gamma + v \tag{8.44}$$

where $z = C'\iota_{T-k}$ is one additional regressor with C defined in (8.34) and ι_{T-k} denoting a vector of ones of dimension $T-k$. In fact, the F-statistic for testing H_0; $\gamma = 0$ turns out to be the square of the Harvey and Collier (1977) t-statistic given in (8.43), see problem 15.

Alternatively, a Sign test may be used to test the null hypothesis of no functional misspecification. Under the null hypothesis, the expected number of positive recursive residuals is equal to $(T-k)/2$. A critical region may therefore be constructed from the binomial distribution. However, Harvey and Collier (1977) suggest that the Sign test tends to lack power compared with the t-test described in (8.43). Nevertheless, it is very simple and it may be more robust to non-normality.

(5) Brown, Durbin and Evans (1975) used recursive residuals to test for structural change over time. The null hypothesis is

$$H_0; \begin{cases} \beta_1 = \beta_2 = .. = \beta_T = \beta \\ \sigma_1^2 = \sigma_2^2 = .. = \sigma_T^2 = \sigma^2 \end{cases} \tag{8.45}$$

where β_t is the vector of coefficients in period t and σ_t^2 is the disturbance variance for that period. The authors suggest a pair of tests. The first is the CUSUM test which computes

$$W_r = \textstyle\sum_{t=k+1}^{r} w_t / s_w \quad \text{for } r = k+1, \ldots, T \tag{8.46}$$

where s_w^2 is an estimate of the variance of the w_t's, given below (8.43). W_r is a cumulative sum and should be plotted against r. Under the null, $E(W_r) = 0$. But, if there is a structural break, W_r will tend to diverge from the horizontal line. The authors suggest checking whether W_r cross a pair of straight lines (see Figure 8.1) which pass through the points $\left\{k, \pm a\sqrt{T-k}\right\}$and $\left\{T, \pm 3a\sqrt{T-k}\right\}$ where a depends upon the chosen significance level α. For example, $a = 0.850, 0.948$, and 1.143 for $\alpha = 10\%, 5\%$, and 1% levels, respectively.

If the coefficients are not constant, there may be a tendency for a disproportionate number of recursive residuals to have the same sign and to push W_r across the boundary. The second test is the cumulative sum of squares (CUSUMSQ) which is based on plotting

$$W_r^* = \textstyle\sum_{t=k+1}^{r} w_t^2 / \sum_{t=k+1}^{T} w_t^2 \quad \text{for } t = k+1, \ldots, T \tag{8.47}$$

against r. Under the null, $E(W_r^*) = (r-k)/(T-k)$ which varies from 0 for $r = k$ to 1 for $r = T$. The significance of the departure of W_r^* from its expected value is assessed by whether W_r^* crosses a pair of lines parallel to $E(W_r^*)$ at a distance c_o above and below this line. Brown, Durbin and Evans (1975) provide values of c_0 for various sample sizes T and levels of significance α.

The CUSUM and CUSUMSQ should be regarded as *data analytic* techniques; i.e., the value of the plots lie in the information to be gained simply by inspecting them. The plots contain more information than can be summarized in a single test statistic. The significance lines constructed are, to paraphrase the authors, best regarded as 'yardsticks' against which to assess the observed plots rather than as formal tests of significance. See Brown et al. (1975) for various examples. Note that the CUSUM and CUSUMSQ are quite general tests for structural change in that they do not require a prior determination of where the structural break takes place. If this is known, the Chow-test will be more powerful. But, if this break is not known, the CUSUM and CUSUMSQ are more appropriate.

Example 2: Table 8.4 reproduces the consumption-income data, over the period 1950-1993, taken from the Economic Report of the President. In addition, the recursive residuals are computed as in (8.30) and exhibited in column 4, starting with 1952 and ending in 1993.

Column 5 gives the CUSUM given by W_r in (8.46) and this is plotted against r in Figure 8.2. The CUSUM crosses the upper 5% line in 1993, showing structural instability in the latter years. This was done using EViews.

The post-sample predictive test for 1993, can be obtained from (8.38) by computing the RSS from 1950-1992 and comparing it with the RSS from 1950-1993. The observed F-statistic is 5.3895 which is distributed as $F(1, 41)$. The same F-statistic is obtained from (8.40) using recursive residuals. In fact,

$$F = w_{1993}^2 / \textstyle\sum_{t=1952}^{1992} w_t^2 / 41 = (339.300414)^2 / 875798.36 \div 41 = 5.3895$$

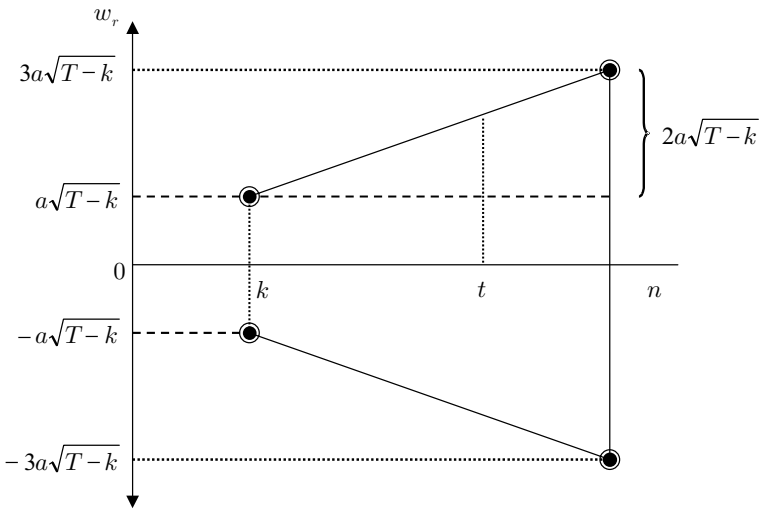


Figure 8.1 CUSUM Critical Values

The Phillips and Harvey (1974) modified von Neuman ratio can be computed from (8.42) and yields an MVNR of 0.4263. The asymptotic distribution of this statistic is $N(2, 4/(44-2))$ which yields a standardized $N(0,1)$ statistic $z = (0.4263-2)/0.3086 = -5.105$. This is highly significant and indicates the presence of positive serial correlation.

The Harvey and Collier (1977) functional misspecification test, given in (8.43), yields $\bar{w} = 66.585$ and $s_w = 140.097$ and an observed t-statistic of 0.0733. This is distributed as t_{41} under the null hypothesis and is not significant.

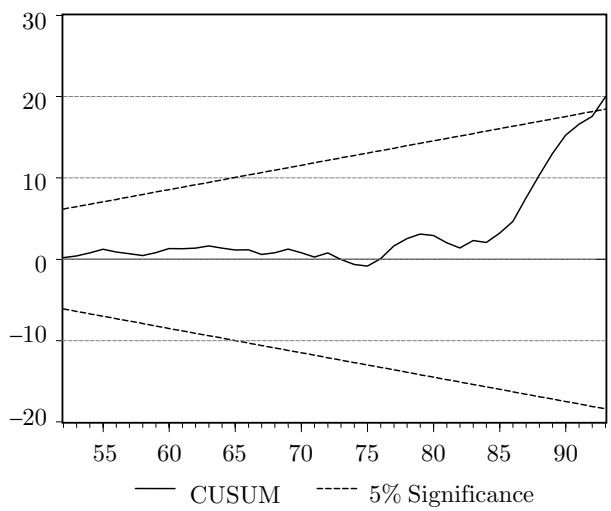


Figure 8.2 CUSUM Plot of Consumption-Income Data

Table 8.4 Consumption-Income Example

YEAR	Consumption	Income	Recursive residuals	Wr of CUSUM
1950	5820	6284	N/A	N/A
1951	5843	6390	N/A	N/A
1952	5917	6476	24.90068062	0.17773916
1953	6054	6640	30.35482733	0.39440960
1954	6099	6628	50.89329055	0.75768203
1955	6365	6879	63.26038894	1.20922986
1956	6440	7080	–49.80590703	0.85371909
1957	6465	7114	–28.40431109	0.65097129
1958	6449	7113	–31.52055902	0.42597994
1959	6658	7256	53.19425584	0.80567648
1960	6698	7264	67.69611375	1.28888618
1961	6740	7382	–2.64655577	1.26999527
1962	6931	7583	9.67914674	1.33908428
1963	7089	7718	39.65882720	1.62216596
1964	7384	8140	–40.12655671	1.33574566
1965	7703	8508	–30.26075567	1.11974669
1966	8005	8822	2.60563312	1.13834551
1967	8163	9114	–78.94146710	0.57486733
1968	8506	9399	27.18506579	0.76891226
1969	8737	9606	64.36319479	1.22833184
1970	8842	9875	–64.90671688	0.76503264
1971	9022	10111	–71.64101261	0.25366456
1972	9425	10414	70.09586688	0.75400351
1973	9752	11013	–113.47532273	–0.05597469
1974	9602	10832	–85.63317118	–0.66721774
1975	9711	10906	–29.42763002	–0.87726992
1976	10121	11192	128.32845919	0.03872884
1977	10425	11406	220.69313327	1.61401959
1978	10744	11851	126.59174925	2.51762185
1979	10876	12039	78.39424741	3.07719401
1980	10746	12005	–25.95557363	2.89192510
1981	10770	12156	–124.17868561	2.00554711
1982	10782	12146	–90.84519334	1.35710105
1983	11179	12349	127.83058064	2.26954599
1984	11617	13029	–30.79462855	2.04973628
1985	12015	13258	159.78087196	3.16024996
1986	12336	13552	201.70712713	4.63001005
1987	12568	13545	405.31056105	7.52308592
1988	12903	13890	390.95384121	10.31368463
1989	13029	14005	373.37091949	12.97877777
1990	13093	14101	316.43123530	15.23743980
1991	12899	14003	188.10968277	16.58015237
1992	13110	14279	134.46128461	17.53992676
1993	13391	14341	339.30041400	19.96182724

8.3 Specification Tests

Specification tests are an important part of model specification in econometrics. In this section, we only study a few of these diagnostic tests. For an excellent summary on this topic, see Wooldridge (2001).

(1) Ramsey's (1969) RESET (Regression Specification Error Test)

Ramsey suggests testing the specification of the linear regression model $y_t = X_t'\beta + u_t$ by augmenting it with a set of regressors Z_t so that the augmented model is

$$y_t = X_t'\beta + Z_t'\gamma + u_t \tag{8.48}$$

If the Z_t's are available then the specification test would reduce to the F-test for H_0; $\gamma = 0$. The crucial issue is the choice of Z_t variables. This depends upon the true functional form under the alternative, which is usually unknown. However, this can be often well approximated by higher powers of the initial regressors, as in the case where the true form is quadratic or cubic. Alternatively, one might approximate it with higher moments of $\hat{y}_t = X_t'\widehat{\beta}_{OLS}$. The popular Ramsey RESET test is carried out as follows:

(1) Regress y_t on X_t and get $\hat{y}_t$.

(2) Regress y_t on X_t, $\hat{y}_t^2$, $\hat{y}_t^3$ and $\hat{y}_t^4$ and test that the coefficients of all the powers of $\hat{y}_t$ are zero. This is an $F_{3,T-k-3}$ under the null.

Note that $\hat{y}_t$ is not included among the regressors because it would be perfectly multicollinear with X_t.[3] Different choices of Z_t's may result in more powerful tests when H_0 is not true. Thursby and Schmidt (1977) carried out an extensive Monte Carlo and concluded that the test based on $Z_t = [X_t^2, X_t^3, X_t^4]$ seems to be generally the best choice.

(2) Utts' (1982) Rainbow Test

The basic idea behind the Rainbow test is that even when the true relationship is nonlinear, a good linear fit can still be obtained over subsets of the sample. The test therefore rejects the null hypothesis of linearity whenever the overall fit is markedly inferior to the fit over a properly selected sub-sample of the data, see Figure 8.3.

Let $e'e$ be the OLS residuals sum of squares from all available n observations and let $\widetilde{e}'\widetilde{e}$ be the OLS residual sum of squares from the middle half of the observations $(T/2)$. Then

$$F = \frac{(e'e - \widetilde{e}'\widetilde{e})/\left(\frac{T}{2}\right)}{\widetilde{e}'\widetilde{e}/\left(\frac{T}{2} - k\right)} \quad \text{is distributed as} \quad F_{\frac{T}{2},\left(\frac{T}{2}-k\right)} \text{ under } H_0 \tag{8.49}$$

Under H_0; $E(e'e/T - k) = \sigma^2 = E\left[\widetilde{e}'\widetilde{e}/\left(\frac{T}{2} - k\right)\right]$, while in general under H_A; $E(e'e/T - k) > E\left[\widetilde{e}'\widetilde{e}/\left(\frac{T}{2} - k\right)\right] > \sigma^2$. The RRSS is $e'e$ because *all* the observations are forced to fit the straight line, whereas the URSS is $\widetilde{e}'\widetilde{e}$ because only a *part* of the observations are forced to fit a straight line. The crucial issue of the Rainbow test is the proper choice of the subsample (the middle $T/2$ observations in case of one regressor). This affects the power of the test and

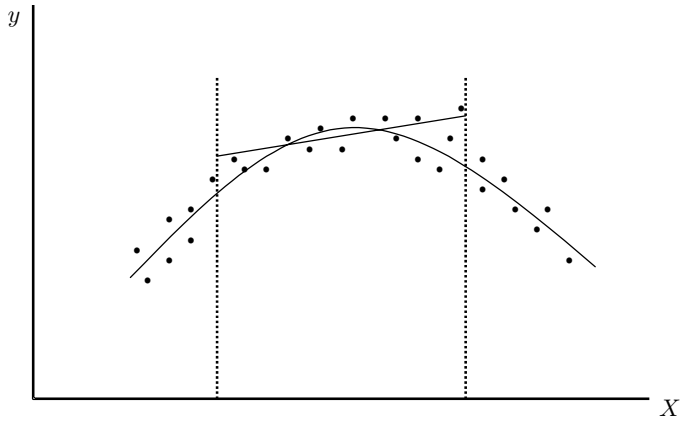


Figure 8.3 The Rainbow Test

not the distribution of the test statistic under the null. Utts (1982) recommends points close to $\bar{X}$, since an incorrect linear fit will in general not be as far off there as it is in the outer region. Closeness to $\bar{X}$ is measured by the magnitude of the corresponding diagonal elements of P_X. Close points are those with low leverage h_{ii}, see section 8.1. The optimal size of the subset depends upon the alternative. Utts recommends about 1/2 of the data points in order to obtain some robustness to outliers. The F-test in (8.49) looks like a Chow test, but differs in the selection of the sub-sample. For example, using the post-sample predictive Chow test, the data are arranged according to time and the first T observations are selected. The Rainbow test arranges the data according to their distance from $\bar{X}$ and selects the first $T/2$ of them.

(3) Plosser, Schwert and White (1982) (PSW) Differencing Test

The differencing test is a general test for misspecification (like Hausman's (1978) test, which will be introduced in the simultaneous equation chapter) but for time-series data only. This test compares OLS and First Difference (FD) estimates of β. Let the differenced model be

$$\dot{y} = \dot{X}\beta + \dot{u} \tag{8.50}$$

where $\dot{y} = Dy$, $\dot{X} = DX$ and $\dot{u} = Du$ where

$$D = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & -1 & \dots & 0 & 0 \\ . & . & . & . & \dots & . & . \\ 0 & 0 & 0 & 0 & \dots & 1 & -1 \end{bmatrix}$$ is the familiar $(T-1) \times T$ differencing matrix.

Wherever there is a constant in the regression, the first column of X becomes zero and is dropped. From (8.50), the FD estimator is given by

$$\widetilde{\beta}_{FD} = (\dot{X}'\dot{X})^{-1}\dot{X}'\dot{y} \tag{8.51}$$

with $\text{var}(\widetilde{\beta}_{FD}) = \sigma^2(\dot{X}'\dot{X})^{-1}\dot{X}'DD'\dot{X}(\dot{X}'\dot{X})^{-1}$ since $\text{var}(\dot{u}) = \sigma^2(DD)'$ and

$$DD' = \begin{bmatrix} 2 & -1 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 \\ : & : & : & : & : & : \\ 0 & 0 & 0 & \dots & 2 & -1 \\ 0 & 0 & 0 & \dots & -1 & 2 \end{bmatrix}$$

The differencing test is based on

$$\widehat{q} = \widetilde{\beta}_{FD} - \widehat{\beta}_{OLS} \quad \text{with} \quad V(\widehat{q}) = \sigma^2[V(\widetilde{\beta}_{FD}) - V(\widehat{\beta}_{OLS})] \tag{8.52}$$

A consistent estimate of $V(\widehat{q})$ is

$$\widehat{V}(\widehat{q}) = \widehat{\sigma}^2 \left[\left(\frac{\dot{X}'\dot{X}}{T} \right)^{-1} \left(\frac{\dot{X}'DD'\dot{X}}{T} \right) \left(\frac{\dot{X}'\dot{X}}{T} \right)^{-1} - \left(\frac{X'X}{T} \right)^{-1} \right] \tag{8.53}$$

where $\widehat{\sigma}^2$ is a consistent estimate of σ^2. Therefore,

$$\Delta = T\widehat{q}'[\widehat{V}(\widehat{q})]^{-1}\widehat{q} \sim \chi^2_k \text{ under } H_0 \tag{8.54}$$

where k is the number of slope parameters if $\widehat{V}(\widehat{q})$ is nonsingular. $\widehat{V}(\widehat{q})$ could be singular, in which case we use a generalized inverse $\widehat{V}^{-}(\widehat{q})$ of $\widehat{V}(\widehat{q})$ and in this case is distributed as χ^2 with degrees of freedom equal to the rank$(\widehat{V}(\widehat{q}))$. This is a special case of the general Hausman (1978) test which will be studied extensively in Chapter 11.

Davidson, Godfrey, and MacKinnon (1985) show that, like the Hausman test, the PSW test is equivalent to a much simpler omitted variables test, the omitted variables being the sum of the lagged and one-period ahead values of the regressors.

Thus if the regression equation we are considering is

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + u_t \tag{8.55}$$

the PSW test involves estimating the expanded regression equation

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \gamma_1 z_{1t} + \gamma_2 z_{2t} + u_t \tag{8.56}$$

where $z_{1t} = x_{1,t+1} + x_{1,t-1}$ and $z_{2t} = x_{2,t+1} + x_{2,t-1}$ and testing the hypothesis $\gamma_1 = \gamma_2 = 0$ by the usual F-test.

If there are lagged dependent variables in the equation, the test needs a minor modification. Suppose that the model is

$$y_t = \beta_1 y_{t-1} + \beta_2 x_t + u_t \tag{8.57}$$

Now the omitted variables would be defined as $z_{1t} = y_t + y_{t-2}$ and $z_{2t} = x_{t+1} + x_{t-1}$. There is no problem with z_{2t} but z_{1t} would be correlated with the error term u_t because of the presence of y_t in it. The solution would be simply to transfer it to the left hand side and write the expanded regression equation in (8.56) as

$$(1 - \gamma_1)y_t = \beta_1 y_{t-1} + \beta_2 x_t + \gamma_1 y_{t-2} + \gamma_2 z_{2t} + u_t \tag{8.58}$$

This equation can be written as

$$y_t = \beta_1^* y_{t-1} + \beta_2^* x_t + \gamma_1^* y_{t-2} + \gamma_2^* z_{2t} + u_t^* \tag{8.59}$$

where all the starred parameters are the corresponding unstarred ones divided by $(1 - \gamma_1)$.

The PSW now tests the hypothesis $\gamma_1^* = \gamma_2^* = 0$. Thus, in the case where the model involves the lagged dependent variable y_{t-1} as an explanatory variable, the only modification needed is that we should use y_{t-2} as the omitted variable, not $(y_t + y_{t-2})$. Note that it is only y_{t-1} that creates a problem, not higher-order lags of y_t, like y_{t-2}, y_{t-3}, and so on. For y_{t-2}, the corresponding z_t will be obtained by adding y_{t-1} to y_{t-3}. This z_t is not correlated with u_t as long as the disturbances are not serially correlated.

(4) Tests for Non-nested Hypothesis

Consider the following two competing non-nested models:

$$H_1;\ y = X_1\beta_1 + \epsilon_1 \tag{8.60}$$

$$H_2;\ y = X_2\beta_2 + \epsilon_2 \tag{8.61}$$

These are non-nested because the explanatory variables under one model are not a subset of the other model even though X_1 and X_2 may share some common variables. In order to test H_1 versus H_2, Cox (1961) modified the LR-test to allow for the non-nested case. The idea behind Cox's approach is to consider to what extent Model I under H_1, is capable of predicting the performance of Model II, under H_2.

Alternatively, one can artificially nest the 2 models

$$H_3;\ y = X_1\beta_1 + X_2^*\beta_2^* + \epsilon_3 \tag{8.62}$$

where X_2^* excludes from X_2 the common variables with X_1. A test for H_1 is simply the F-test for $H_0;\ \beta_2^* = 0$.

Criticism: This tests H_1 versus H_3 which is a (Hybrid) of H_1 and H_2 and not H_1 versus H_2.

Davidson and MacKinnon (1981) proposed (testing $\alpha = 0$) in the linear combination of H_1 and H_2:

$$y = (1 - \alpha)X_1\beta_1 + \alpha X_2\beta_2 + \epsilon \tag{8.63}$$

where α is an unknown scalar. Since α is not identified, we replace β_2 by $\widehat{\beta}_{2,OLS} = (X_2'X_2/T)^{-1}$ $(X_2'y/T)$ the regression coefficient estimate obtained from running y on X_2 under H_2, i.e., (1) Run y on X_2 get $\hat{y}_2 = X_2\widehat{\beta}_{2,OLS}$; (2) Run y on X_1 and $\widehat{y}_2$ and test that the coefficient of $\widehat{y}_2$ is zero. This is known as the J-test and this is asymptotically $N(0,1)$ under H_1.

Fisher and McAleer (1981) suggested a modification of the J-test known as the JA test.

$$\text{Under } H_1; \text{plim}\widehat{\beta}_2 = \text{plim}(X_2'X_2/T)^{-1}\text{plim}(X_2'X_1/T)\beta_1 + 0 \tag{8.64}$$

Therefore, they propose replacing $\widehat{\beta}_2$ by $\widetilde{\beta}_2 = (X_2'X_2)^{-1}(X_2'X_1)\widehat{\beta}_{1,OLS}$ where $\widehat{\beta}_{1,OLS} = (X_1'X_1)^{-1}$ $X_1'y$. The steps for the JA-test are as follows:

1. Run y on X_1 get $\widehat{y}_1 = X_1\widehat{\beta}_{1,OLS}$.
2. Run $\widehat{y}_1$ on X_2 get $\widetilde{y}_2 = X_2(X_2'X_2)^{-1}X_2'\widehat{y}_1$.
3. Run y on X_1 and $\widetilde{y}_2$ and test that the coefficient of $\widetilde{y}_2$ is zero. This is the simple t-statistic on the coefficient of $\widetilde{y}_2$. The J and JA tests are asymptotically equivalent.

Criticism: Note the asymmetry of H_1 and H_2. Therefore one should reverse the role of these hypotheses and test again.

In this case one can get the four scenarios depicted in Table 8.5. In case both hypotheses are not rejected, the data are not rich enough to discriminate between the two hypotheses. In case both hypotheses are rejected neither model is useful in explaining the variation in y. In case one hypothesis is rejected while the other is not, one should remember that the non-rejected hypothesis may still be brought down by another challenger hypothesis.

Small Sample Properties: (i) The J-test tends to reject the null more frequently than it should. Also, the JA test has relatively low power when K_1, the number of parameters in H_1 is larger than K_2, the number of parameters in H_2. Therefore, one should use the JA test when K_1 is about the same size as K_2, i.e., the same number of non-overlapping variables. (ii) If both H_1 and H_2 are false, these tests are inferior to the standard diagnostic tests. In practice, use higher significance levels for the J-test, and supplement it with the artificially nested F-test and standard diagnostic tests.

Table 8.5 Non-nested Hypothesis Testing

		$\alpha = 0$	
		Not Rejected	Rejected
$\alpha = 1$	Not Rejected	Both H_1 and H_2 are not rejected	H_1 rejected H_2 not rejected
	Rejected	H_1 not rejected H_2 rejected	Both H_1 and H_2 are rejected

Note: J and JA tests are one degree of freedom tests, whereas the artificially nested F-test is not.

For a recent summary of non-nested hypothesis testing, see Pesaran and Weeks (2001). Examples of non-nested hypothesis encountered in empirical economic research include linear versus log-linear models, see section 8.5. Also, logit versus probit models in discrete choice, see Chapter 13 and exponential versus Weibull distributions in the analysis of duration data. In the logit versus probit specification, the set of regressors is most likely to be the same. It is only the form of the distribution functions that separate the two models. Pesaran and Weeks (2001, p. 287) emphasize the differences between hypothesis testing and model selection:

> *The model selection process treats all models under consideration symmetrically, while hypothesis testing attributes a different status to the null and to the alternative hypotheses and by design treats the models asymmetrically. Model selection always*

ends in a definite outcome, namely one of the models under consideration is selected for use in decision making. Hypothesis testing on the other hand asks whether there is any statistically significant evidence (in the Neyman-Pearson sense) of departure from the null hypothesis in the direction of one or more alternative hypotheses. Rejection of the null hypothesis does not necessarily imply acceptance of any one of the alternative hypotheses; it only warns the investigator of possible shortcomings of the null that is being advocated. Hypothesis testing does not seek a definite outcome and if carried out with due care need not lead to a favorite model. For example, in the case of nonnested hypothesis testing it is possible for all models under consideration to be rejected, or all models to be deemed as observationally equivalent.

They conclude that the choice between hypothesis testing and model selection depends on the primary objective of one's study. Model selection may be more appropriate when the objective is decision making, while hypothesis testing is better suited to inferential problems.

A model may be empirically adequate for a particular purpose, but of little relevance for another use... In the real world where the truth is elusive and unknowable both approaches to model evaluation are worth pursuing.

(5) White's (1982) Information-Matrix (IM) Test

This is a general specification test much like the Hausman (1978) specification test which will be considered in details in Chapter 11. The latter is based on two different estimates of the regression coefficients, while the former is based on two different estimates of the Information Matrix I(θ) where $\theta' = (\beta', \sigma^2)$ in the case of the linear regression studied in Chapter 7. The first estimate of $I(\theta)$ evaluates the expectation of the second derivatives of the log-likelihood at the MLE, i.e., $-E(\partial^2 \log L/\partial\theta\partial\theta')$ at $\widehat{\theta}_{mle}$ while the second sum up the outer products of the score vectors $\sum_{i=1}^{n}(\partial \log L_i(\theta)/\partial\theta)(\partial \log L_i(\theta)/\partial\theta)'$ evaluated at $\widehat{\theta}_{mle}$. This is based on the fundamental identity that

$$I(\theta) = -E(\partial^2 \log L/\partial\theta\partial\theta') = E(\partial \log L/\partial\theta)(\partial \log L/\partial\theta)'$$

If the model estimated by MLE is not correctly specified, this equality will not hold. From Chapter 7, equation (7.19), we know that for the linear regression model with normal disturbances, the first estimate of $I(\theta)$ denoted by $I_1(\widehat{\theta}_{mle})$ is given by

$$I_1(\widehat{\theta}_{MLE}) = \begin{bmatrix} X'X/\widehat{\sigma}^2 & 0 \\ 0 & n/2\widehat{\sigma}^4 \end{bmatrix} \tag{8.65}$$

where $\widehat{\sigma}^2 = e'e/n$ is the MLE of σ^2 and e denotes the OLS residuals.

Similarly, one can show that the second estimate of $I(\theta)$ denoted by $I_2(\theta)$ is given by

$$\begin{aligned} I_2(\theta) &= \left(\frac{\sum_{i=1}^{n}(\partial \log L_i(\theta)}{\partial\theta}\right)\left(\frac{(\partial \log L_i(\theta)}{\partial\theta}\right)' \\ &= \sum_{i=1}^{n}\begin{bmatrix} \dfrac{u_i^2 x_i x_i'}{\sigma^4} & \dfrac{-u_i x_i}{2\sigma^4} + \dfrac{u_i^3 x_i}{2\sigma^6} \\ -\dfrac{u_i x_i'}{2\sigma^4} + \dfrac{u_i^3 x_i'}{2\sigma^6} & \dfrac{1}{4\sigma^4} - \dfrac{u_i^2}{2\sigma^6} + \dfrac{u_i^4}{4\sigma^8} \end{bmatrix} \end{aligned} \tag{8.66}$$

where x_i is the i-th row of X. Substituting the MLE we get

$$I_2(\widehat{\theta}_{MLE}) = \begin{bmatrix} \frac{\sum_{i=1}^n e_i^2 x_i x_i'}{\widehat{\sigma}^4} & \frac{\sum_{i=1}^n e_i^3 x_i}{2\widehat{\sigma}^6} \\ \frac{\sum_{i=1}^n e_i^3 x_i'}{2\widehat{\sigma}^6} & -\frac{n}{4\widehat{\sigma}^4} + \frac{\sum_{i=1}^n e_i^4}{4\widehat{\sigma}^8} \end{bmatrix} \tag{8.67}$$

where we used the fact that $\sum_{i=1}^n e_i x_i = 0$. If the model is correctly specified and the disturbances are normal then

$$\text{plim } I_1(\widehat{\theta}_{MLE})/n = \text{plim } I_2(\widehat{\theta}_{MLE})/n = I(\theta)$$

Therefore, the Information Matrix (IM) test rejects the model when

$$[I_2(\widehat{\theta}_{MLE}) - I_1(\widehat{\theta}_{MLE})]/n \tag{8.68}$$

is too large. These are two matrices with $(k+1)$ by $(k+1)$ elements since β is $k \times 1$ and σ^2 is a scalar. However, due to symmetry, this reduces to $(k+2)(k+1)/2$ unique elements. Hall (1987) noted that the first $k(k+1)/2$ unique elements obtained from the first $k \times k$ block of (8.68) have a typical element $\sum_{i=1}^n (e_i^2 - \widehat{\sigma}^2) x_{ir} x_{is}/n\widehat{\sigma}^4$ where r and s denote the r-th and s-th explanatory variables with $r, s = 1, 2, \ldots, k$. This term measures the discrepancy between the OLS estimates of the variance-covariance matrix of $\widehat{\beta}_{OLS}$ and its robust counterpart suggested by White (1980), see Chapter 5. The next k unique elements correspond to the off-diagonal block $\sum_{i=1}^n e_i^3 x_i/2n\widehat{\sigma}^6$and this measures the discrepancy between the estimates of the $\text{cov}(\widehat{\beta}, \widehat{\sigma}^2)$. The last element correspond to the difference in the bottom right elements, i.e., the two estimates of $\widehat{\sigma}^2$. This is given by

$$\left[-\frac{3}{4\widehat{\sigma}^4} + \frac{1}{n} \sum_{i=1}^n e_i^4 / 4\widehat{\sigma}^8 \right]$$

These $(k+1)(k+2)/2$ unique elements can be arranged in vector form $D(\theta)$ which has a limiting normal distribution with zero mean and some covariance matrix $V(\theta)$ under the null. One can show, see Hall (1987) or Krämer and Sonnberger (1986) that if $V(\theta)$ is estimated from the sample moments of these terms, that the IM test statistic is given by

$$m = nD'(\theta)[V(\theta)]^{-1} D(\theta) \overset{H_0}{\rightarrow} \chi^2_{(k+1)(k+2)/2} \tag{8.69}$$

In fact, Hall (1987) shows that this statistic is the sum of three asymptotically independent terms

$$m = m_1 + m_2 + m_3 \tag{8.70}$$

where $m_1 =$ a particular version of White's heteroskedasticity test; $m_2 = n$ times the explained sum of squares from the regression of e_i^3 on x_i divided by $6\widehat{\sigma}^6$; and

$$m_3 = \frac{n}{24\widehat{\sigma}^8} \left(\sum_{i=1}^n e_i^4/n - 3\widehat{\sigma}^4 \right)^2$$

which is similar to the Jarque-Bera test for normality of the disturbances given in Chapter 5.

It is clear that the IM test will have power whenever the disturbances are non-normal or heteroskedastic. However, Davidson and MacKinnon (1992) demonstrated that the IM test

considered above will tend to reject the model when true, much too often, in finite samples. This problem gets worse as the number of degrees of freedom gets large. In Monte Carlo experiments, Davidson and MacKinnon (1992) showed that for a linear regression model with ten regressors, the IM test rejected the null at the 5% level, 99.9% of the time for $n = 200$. This problem did not disappear when n increased. In fact, for $n = 1000$, the IM test still rejected the null 92.7% of the time at the 5% level.

These results suggest that it may be more useful to run individual tests for non-normality, heteroskedasticity and other misspecification tests considered above rather than run the IM test. These tests may be more powerful and more informative than the IM test. Alternative methods of calculating the IM test with better finite-sample properties are suggested in Orme (1990), Chesher and Spady (1991) and Davidson and MacKinnon (1992).

Example 3: For the consumption-income data given in Table 5.1, we first compute the RESET test from the consumption-income regression given in Chapter 5. Using EViews, one clicks on *stability tests* and then selects RESET. You will be prompted with the option of the number of fitted terms to include (i.e., powers of $\widehat{y}$). Table 8.6 shows the RESET test including $\widehat{y}^2$ and $\widehat{y}^3$. The F-statistic for their joint-significance is equal to 40.98. This is significant and indicates misspecification.

Next, we compute Utts (1982) Rainbow test. Table 8.7 gives the middle half of the sample, i.e., 1962-1983, and the SAS regression using this data. The RSS of these middle observations is given by $\widetilde{e}'\widetilde{e} = 178245.867$, while the RSS for the entire sample is given by $e'e = 990923.131$ so that the observed F-statistic given in (8.49) can be computed as follows:

$$F = \frac{(990923.131 - 178245.867)/22}{178245.867/20} = 4.145$$

Table 8.6 Ramsey RESET Test

F-statistic	40.97877	Probability		0.00000
Log likelihood ratio	49.05091	Probability		0.00000
Test Equation:				
Dependent Variable:	CONSUM			
Method:	Least Squares			
Sample:	1950 1993			
Included observations:	44			
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	–3210.837	1299.234	–2.471330	0.0178
Y	2.079184	0.395475	5.257428	0.0000
FITTED^2	–0.000165	4.71E–05	–3.494615	0.0012
FITTED^3	6.66E–09	1.66E–09	4.012651	0.0003
R-squared	0.998776	Mean dependent var		9250.54
Adjusted R-squared	0.998684	S.D. dependent var		2484.62
S.E. of regression	90.13960	Akaike info criterion		11.9271
Sum squared resid	325005.9	Schwarz criterion		12.0893
Log likelihood	–258.3963	F-statistic		10876.9
Durbin-Watson stat	1.458024	Prob (F-statistic)		0.00000

Table 8.7 Utts (1982) Rainbow Test

OBS	YEAR	INCOME	CONSUMP	LEVERAGE
1	1962	7583	6931	0.0440
2	1963	7718	7089	0.0419
3	1964	8140	7384	0.0359
4	1965	8508	7703	0.0315
5	1966	8822	8005	0.0285
6	1967	9114	8163	0.0263
7	1968	9399	8506	0.0246
8	1969	9606	8737	0.0238
9	1970	9875	8842	0.0230
10	1971	10111	9022	0.0227
11	1972	10414	9425	0.0229
12	1973	11013	9752	0.0250
13	1974	10832	9602	0.0241
14	1975	10906	9711	0.0244
15	1976	11192	10121	0.0260
16	1977	11406	10425	0.0275
17	1978	11851	10744	0.0316
18	1979	12039	10876	0.0337
19	1980	12005	10746	0.0333
20	1981	12156	10770	0.0352
21	1982	12146	10782	0.0350
22	1983	12349	11179	0.0377

Dependent Variable: CONSUMPTION

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	37715232.724	37715232.724	4231.821	0.0001
Error	20	178245.86703	8912.29335		
C Total	21	37893478.591			

Root MSE	94.40494	R-square	0.9953
Dep Mean	9296.13636	Adj R-sq	0.9951
C.V.	1.01553		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	290.338091	139.89449574	2.075	0.0511
INCOME	1	0.872098	0.01340607	65.052	0.0001

Table 8.8 PSW Differencing Test

Dependent Variable:	CONSUM			
Method:	Least Squares			
Sample (adjusted):	1951 1993			
Included observations:	43 after adjusting endpoints			
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	–60.28744	94.40916	–0.638576	0.5267
Y	0.501803	0.273777	1.832891	0.0743
Z	0.208394	0.137219	1.518701	0.1367
R-squared	0.996373	Mean dependent var		9330.326
Adjusted R-squared	0.996191	S.D. dependent var		2456.343
S.E. of regression	151.5889	Akaike info criterion		12.94744
Sum squared resid	919167.2	Schwarz criterion		13.07031
Log likelihood	–275.3699	F-statistic		5493.949
Durbin-Watson stat	0.456301	Prob (F-statistic)		0.00000

This is distributed as $F_{22,20}$ under the null hypothesis and rejects the hypothesis of linearity.

The PSW differencing test, given in (8.54), is based upon $\widehat{\beta}_{OLS} = 0.915623$ with $V(\widehat{\beta}_{OLS}) = 0.0000747926$, and $\widetilde{\beta}_{FD} = 0.843298$ with $V(\widetilde{\beta}_{FD}) = 0.00732539$. Therefore, $\widehat{q} = (0.843298 - 0.915623) = -0.072325$ with $\widehat{V}(\widehat{q})/T = 0.00732539 - 0.0000747926 = 0.0072506$ and $\Delta = T\widehat{q}'[\widehat{V}(\widehat{q})]^{-1}\widehat{q} = 0.721$ which is distributed as χ^2_1 under the null. This is not significant and does not reject the null of no misspecification. The artificial regression given in (8.56) with $Z_t = Y_t + Y_{t-1}$ is given in Table 8.8. The t-statistic for Z_t is 1.52 and has a p-value of 0.137 which is not significant.

Now consider the two competing non-nested models:

$$H_1;\ C_t = \beta_0 + \beta_1 Y_t + \beta_2 Y_{t-1} + u_t \quad H_2;\ C_t = \gamma_0 + \gamma_1 Y_t + \gamma_2 C_{t-1} + v_t$$

The two non-nested models share Y_t as a common variable. The artificial model that nests these two models is given by:

$$H_3;\ C_t = \delta_0 + \delta_1 Y_t + \delta_2 Y_{t-1} + \delta_3 C_{t-1} + \epsilon_t$$

Table 8.9, runs regression (1) given by H_2 and obtains the predicted values $\widehat{C}_2(C2HAT)$. Regression (2) runs consumption on a constant, income, lagged income and $C2HAT$. The coefficient of this last variable is 1.64 and is statistically significant with a t-value of 7.80. This is the Davidson and MacKinnon (1981) J-test. In this case, H_1 is rejected but H_2 is not rejected. The JA-test, given by Fisher and McAleer (1981) runs the regression in H_1 and keeps the predicted values $\widehat{C}_1(C1HAT)$. This is done in regression (3). Then $C1HAT$ is run on a constant, income and lagged consumption and the predicted values are stored as $\widetilde{C}_2(C2TILDE)$. This is done in regression (5). The last step runs consumption on a constant, income, lagged income and $C2TILDE$, see regression (6). The coefficient of this last variable is 6.53 and is statistically significant with a t-value of 7.80. Again H_1 is rejected but H_2 is not rejected.

Reversing the roles of H_1 and H_2, the J and JA-tests are repeated. In fact, regression (4) runs consumption on a constant, income, lagged consumption and $\widehat{C}_1$ (which was obtained from

regression (3)). The coefficient on $\widehat{C}_1$ is -2.54 and is statistically significant with a t-value of -4.13. This J-test rejects H_2 but does not reject H_1. Regression (7) runs $\widehat{C}_2$ on a constant, income and lagged income and the predicted values are stored as $\widetilde{C}_1$(C1TILDE). The last step of the JA test runs consumption on a constant, income, lagged consumption and $\widetilde{C}_1$, see regression (8). The coefficient of this last variable is -1.18 and is statistically significant with a t-value of -4.13. This JA test rejects H_2 but not H_1. The artificial model, given in H_3, is also estimated, see regression (9). One can easily check that the corresponding F-tests reject H_1 against H_3 and also H_2 against H_3. In sum, all evidence indicates that both C_{t-1} and Y_{t-1} are important to include along with Y_t. Of course, the true model is not known and could include higher lags of both Y_t and C_t.

8.4 Nonlinear Least Squares and the Gauss-Newton Regression[4]

So far we have been dealing with linear regressions. But, in reality, one might face a nonlinear regression of the form:

$$y_t = x_t(\beta) + u_t \quad \text{for } t = 1, 2, \ldots, T \tag{8.71}$$

where $u_t \sim \text{IID}(0, \sigma^2)$ and $x_t(\beta)$ is a scalar nonlinear regression function of k unknown parameters β. It can be interpreted as the expected value of y_t conditional on the values of the independent variables. Nonlinear least squares minimizes $\sum_{t=1}^{T}(y_t - x_t(\beta))^2 = (y - x(\beta))'(y - x(\beta))$. The first-order conditions for minimization yield

$$X'(\widehat{\beta})(y - x(\widehat{\beta})) = 0 \tag{8.72}$$

where $X(\beta)$ is a $T \times k$ matrix with typical element $X_{tj}(\beta) = \partial x_t(\beta)/\partial \beta_j$ for $j = 1, \ldots, k$. The solution to these k equations yield the Nonlinear Least Squares (NLS) estimates of β denoted by $\widehat{\beta}_{NLS}$. These normal equations given in (8.72) are similar to those in the linear case in that they require the vector of residuals $y - x(\widehat{\beta})$ to be orthogonal to the matrix of derivatives $X(\widehat{\beta})$. In the linear case, $x(\widehat{\beta}) = X\widehat{\beta}_{OLS}$ and $X(\widehat{\beta}) = X$ where the latter is independent of $\widehat{\beta}$. Because of this dependence of the fitted values $x(\widehat{\beta})$ as well as the matrix of derivatives $X(\widehat{\beta})$ on $\widehat{\beta}$, one in general cannot get explicit analytical solution to these NLS first-order equations. Under fairly general conditions, see Davidson and MacKinnon (1993), one can show that the $\widehat{\beta}_{NLS}$ has asymptotically a normal distribution with mean β_0 and asymptotic variance $\sigma_0^2(X'(\beta_0)X(\beta_0))^{-1}$, where β_0 and σ_0 are the true values of the parameters generating the data. Similarly, defining

$$s^2 = (y - x(\widehat{\beta}_{NLS}))'(y - x(\widehat{\beta}_{NLS}))/(T - k)$$

we get a feasible estimate of this covariance matrix as $s^2(X'(\widehat{\beta})X(\widehat{\beta}))^{-1}$. If the disturbances are normally distributed then NLS is MLE and therefore asymptotically efficient as long as the model is correctly specified, see Chapter 7.

Taking the first-order Taylor series approximation around some arbitrary parameter vector β^*, we get

$$y = x(\beta^*) + X(\beta^*)(\beta - \beta^*) + \text{higher-order terms} + u \tag{8.73}$$

or

$$y - x(\beta^*) = X(\beta^*)b + \text{residuals} \tag{8.74}$$

Table 8.9 Non-nested J and JA Test

REGRESSION 1

Dependent Variable: C
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	252896668.15	126448334.08	9813.926	0.0001
Error	40	515383.28727	12884.58218		
C Total	42	253412051.44			

Root MSE	113.51027	R-square	0.9980
Dep Mean	9330.32558	Adj R-sq	0.9979
C.V.	1.21657		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	–17.498706	70.43957345	–0.248	0.8051
Y	1	0.475030	0.07458032	6.369	0.0001
CLAG	1	0.488458	0.08203649	5.954	0.0001

REGRESSION 2

Dependent Variable: C
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	253053288.45	84351096.151	9169.543	0.0001
Error	39	358762.98953	9199.05101		
C Total	42	253412051.44			

Root MSE	95.91168	R-square	0.9986
Dep Mean	9330.32558	Adj R-sq	0.9985
C.V.	1.02796		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	–12.132250	60.05133697	–0.202	0.8409
Y	1	–0.058489	0.13107211	–0.446	0.6579
YLAG	1	–0.529709	0.12837635	–4.126	0.0002
C2HAT	1	1.637806	0.20983763	7.805	0.0001

Table 8.9 (continued)

REGRESSION 3

Dependent Variable: C
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	252492884.22	126246442.11	5493.949	0.0001
Error	40	919167.21762	22979.18044		
C Total	42	253412051.44			

Root MSE	151.58885	R-square	0.9964
Dep Mean	9330.32558	Adj R-sq	0.9962
C.V.	1.62469		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	–60.287436	94.40916038	–0.639	0.5267
Y	1	0.710197	0.13669864	5.195	0.0001
YLAG	1	0.208394	0.13721871	1.519	0.1367

REGRESSION 4

Dependent Variable: C
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	253053288.45	84351096.151	9169.543	0.0001
Error	39	358762.98953	9199.05101		
C Total	42	253412051.44			

Root MSE	95.91168	R-square	0.9986
Dep Mean	9330.32558	Adj R-sq	0.9985
C.V.	1.02796		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	–194.034085	73.30021926	–2.647	0.0117
Y	1	2.524742	0.50073386	5.042	0.0001
CLAG	1	0.800000	0.10249693	7.805	0.0001
C1HAT	1	–2.541862	0.61602660	–4.126	0.0002

Table 8.9 (continued)

Regression 5

Dependent Variable: C1HAT
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	252468643.58	126234321.79	208301.952	0.0001
Error	40	24240.64119	606.01603		
C Total	42	252492884.22			

Root MSE	24.61739	R-square	0.9999
Dep Mean	9330.32558	Adj R-sq	0.9999
C.V.	0.26384		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	–69.451205	15.27649082	–4.546	0.0001
Y	1	0.806382	0.01617451	49.855	0.0001
CLAG	1	0.122564	0.01779156	6.889	0.0001

Regression 6

Dependent Variable: C
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	253053288.45	84351096.151	9169.543	0.0001
Error	39	358762.98953	9199.05101		
C Total	42	253412051.44			

Root MSE	95.91168	R-square	0.9986
Dep Mean	9330.32558	Adj R-sq	0.9985
C.V.	1.02796		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	412.528849	85.07504557	4.849	0.0001
Y	1	–4.543882	0.67869220	–6.695	0.0001
YLAG	1	–0.529709	0.12837635	–4.126	0.0002
C2TILDE	1	6.527181	0.83626992	7.805	0.0001

Table 8.9 (continued)

REGRESSION 7

Dependent Variable: C2HAT
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	252687749.95	126343874.98	24190.114	0.0001
Error	40	208918.20026	5222.95501		
C Total	42	252896668.15			

Root MSE	72.27001	R-square	0.9992
Dep Mean	9330.32558	Adj R-sq	0.9991
C.V.	0.77457		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	–29.402245	45.00958513	–0.653	0.5173
Y	1	0.469339	0.06517110	7.202	0.0001
YLAG	1	0.450666	0.06541905	6.889	0.0001

REGRESSION 8

Dependent Variable: C
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	253053288.45	84351096.151	9169.543	0.0001
Error	39	358762.98953	9199.05101		
C Total	42	253412051.44			

Root MSE	95.91168	R-square	0.9986
Dep Mean	9330.32558	Adj R-sq	0.9985
C.V.	1.02796		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	–75.350919	61.14775168	–1.232	0.2252
Y	1	1.271176	0.20297803	6.263	0.0001
CLAG	1	0.800000	0.10249693	7.805	0.0001
C1TILDE	1	–1.175392	0.28485929	–4.126	0.0002

Table 8.9 (continued)

Regression 9

Dependent Variable: C
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	253053288.45	84351096.151	9169.543	0.0001
Error	39	358762.98953	9199.05101		
C Total	42	253412051.44			

Root MSE	95.91168	R-square	0.9986
Dep Mean	9330.32558	Adj R-sq	0.9985
C.V.	1.02796		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	–40.791743	59.78575886	–0.682	0.4991
Y	1	0.719519	0.08649875	8.318	0.0001
YLAG	1	–0.529709	0.12837635	–4.126	0.0002
CLAG	1	0.800000	0.10249693	7.805	0.0001

This is the simplest version of the Gauss-Newton Regression, see Davidson and MacKinnon (1993). In this case the higher-order terms and the error term are combined in the residuals and $(\beta - \beta^*)$ is replaced by b, a parameter vector that can be estimated. If the model is linear, $X(\beta^*)$ is the matrix of regressors X and the GNR regresses a residual on X. If $\beta^*=\widehat{\beta}_{NLS}$, the unrestricted NLS estimator of β, then the GNR becomes

$$y - \widehat{x} = \widehat{X}b + \text{residuals} \tag{8.75}$$

where $\widehat{x} \equiv x(\widehat{\beta}_{NLS})$ and $\widehat{X} \equiv X(\widehat{\beta}_{NLS})$. From the first-order conditions of NLS we get $(y - \widehat{x})'\widehat{X} = 0$. In this case, OLS on this GNR yields $\widehat{b}_{OLS} = (\widehat{X}'\widehat{X})^{-1}\widehat{X}'(y - \widehat{x}) = 0$ and this GNR has no explanatory power. However, this regression can be used to (i) check that the first-order conditions given in (8.72) are satisfied. For example, one could check that the t-statistics are of the 10^{-3} order, and that R^2 is zero up to several decimal places; (ii) compute estimated covariance matrices. In fact, this GNR prints out $s^2(\widehat{X}'\widehat{X})^{-1}$, where $s^2 = (y-\widehat{x})'(y-\widehat{x})/(T-k)$ is the OLS estimate of the regression variance. This can be verified easily using the fact that this GNR has no explanatory power. This method of computing the estimated variance-covariance matrix is useful especially in cases where $\widehat{\beta}$ has been obtained by some method other than NLS. For example, sometimes the model is nonlinear only in one or two parameters which are known to be in a finite range, say between zero and one. One can then search over this range, running OLS regressions and minimizing the residual sum of squares. This search procedure can be repeated over finer grids to get more accuracy. Once the final parameter estimate is found, one can run the GNR to get estimates of the variance-covariance matrix.

Testing Restrictions (GNR Based on the Restricted NLS Estimates)

The best known use for the GNR is to test restrictions. These are based on the LM principle which requires only the restricted estimator. In particular, consider the following competing hypotheses:

$$H_0; y = x(\beta_1, 0) + u \qquad H_1; y = x(\beta_1, \beta_2) + u$$

where $u \sim \text{IID}(0, \sigma^2 I)$ and β_1 and β_2 are $k \times 1$ and $r \times 1$, respectively. Denote by $\widetilde{\beta}$ the restricted NLS estimator of β, in this case $\widetilde{\beta}' = (\widetilde{\beta}_1', 0)$.

The GNR evaluated at this restricted NLS estimator of β is

$$(y - \widetilde{x}) = \widetilde{X}_1 b_1 + \widetilde{X}_2 b_2 + \text{residuals} \tag{8.76}$$

where $\widetilde{x} = x(\widetilde{\beta})$ and $\widetilde{X}_i = X_i(\widetilde{\beta})$ with $X_i(\beta) = \partial x / \partial \beta_i$ for $i = 1, 2$.

By the FWL Theorem this yields the same estimate of b_2 as

$$\bar{P}_{\widetilde{X}_1}(y - \widetilde{x}) = \bar{P}_{\widetilde{X}_1}\widetilde{X}_2 b_2 + \text{residuals} \tag{8.77}$$

But $\bar{P}_{\widetilde{X}_1}(y - \widetilde{x}) = (y - \widetilde{x}) - P_{\widetilde{X}_1}(y - \widetilde{x}) = (y - \widetilde{x})$ since $\widetilde{X}_1'(y - \widetilde{x}) = 0$ from the first-order conditions of restricted NLS. Hence, (8.77) reduces to

$$(y - \widetilde{x}) = \bar{P}_{\widetilde{X}_1}\widetilde{X}_2 b_2 + \text{residuals} \tag{8.78}$$

Therefore,

$$b_{2,OLS} = (\widetilde{X}_2'\bar{P}_{\widetilde{X}_1}\widetilde{X}_2)^{-1}\widetilde{X}_2'\bar{P}_{\widetilde{X}_1}(y - \widetilde{x}) = (\widetilde{X}_2'\bar{P}_{\widetilde{X}_1}\widetilde{X}_2)^{-1}\widetilde{X}_2'(y - \widetilde{x}) \tag{8.79}$$

and the residual sums of squares is $(y - \widetilde{x})'(y - \widetilde{x}) - (y - \widetilde{x})'\widetilde{X}_2(\widetilde{X}_2'\bar{P}_{\widetilde{X}_1}\widetilde{X}_2)^{-1}\widetilde{X}_2'(y - \widetilde{x})$.

If $\widetilde{X}_2$ was excluded from the regression in (8.76), $(y - \widetilde{x})'(y - \widetilde{x})$ would be the residual sum of squares. Therefore, the reduction in the residual sum of squares brought about by the inclusion of $\widetilde{X}_2$ is

$$(y - \widetilde{x})'\widetilde{X}_2(\widetilde{X}_2'\bar{P}_{\widetilde{X}_1}\widetilde{X}_2)^{-1}\widetilde{X}_2'(y - \widetilde{x})$$

This is also equal to the explained sum of squares from (8.76) since $\widetilde{X}_1$ has no explanatory power. This sum of squares divided by a consistent estimate of σ^2 is asymptotically distributed as χ_r^2 under the null.

Different consistent estimates of σ^2 yield different test statistics. The two most common test statistics for H_0 based on this regression are the following: (1) TR_u^2 where R_u^2 is the *uncentered* R^2 of (8.76) and (2) the F-statistic for $b_2 = 0$. The first statistic is given by $TR_u^2 = T(y - \widetilde{x})'\widetilde{X}_2(\widetilde{X}_2'\bar{P}_{\widetilde{X}_1}\widetilde{X}_2)^{-1}\widetilde{X}_2'(y - \widetilde{x})/(y - \widetilde{x})'(y - \widetilde{x})$ where the *uncentered* R^2 was defined in the Appendix to Chapter 3. This statistic implicitly divides the explained sum of squares term by $\widetilde{\sigma}^2 = (\text{restricted residual sums of squares})/T$. This is equivalent to the LM-statistic obtained by running the artificial regression $(y - \widetilde{x})/\widetilde{\sigma}$ on $\widetilde{X}$ and getting the explained sum of squares. Regression packages print the *centered* R^2. This is equal to the *uncentered* R_u^2 as long as there is a constant in the restricted regression so that $(y - \widetilde{x})$ sum to zero.

The F-statistic for $b_2 = 0$ from (8.76) is

$$\frac{(RRSS - URSS)/r}{URSS/(T-k)} = \frac{(y-\widetilde{x})'\widetilde{X}_2(\widetilde{X}_2'\bar{P}_{\widetilde{X}_1}\widetilde{X}_2)^{-1}\widetilde{X}_2'(y-\widetilde{x})/r}{[(y-\widetilde{x})'(y-\widetilde{x}) - (y-\widetilde{x})'\widetilde{X}_2(\widetilde{X}_2'\bar{P}_{\widetilde{X}_1}\widetilde{X}_2)^{-1}\widetilde{X}_2'(y-\widetilde{x})]/(T-k)} \tag{8.80}$$

The denominator is the OLS estimate of σ^2 from (8.76) which tends to σ_0^2 as $T \to \infty$. Hence (rF-statistic $\to \chi_r^2$). In small samples, use the F-statistic.

Diagnostic Tests for Linear Regression Models

Variable addition tests suggested by Pagan and Hall (1983) consider the additional variables Z of dimension $(T \times r)$ and test whether their coefficients are zero using an F-test from the regression

$$y = X\beta + Z\gamma + u \tag{8.81}$$

If H_0; $\gamma = 0$ is true, the model is $y = X\beta + u$ and there is no misspecification. The GNR for this restriction would run the following regression:

$$\bar{P}_X y = Xb + Zc + \text{residuals} \tag{8.82}$$

and test that c is zero. By the FWL Theorem, (8.82) yields the same residual sum of squares as

$$\bar{P}_X y = \bar{P}_X Zc + \text{residuals} \tag{8.83}$$

Applying the FWL Theorem to (8.81) we get the same residual sum of squares as the regression in (8.83). The F-statistic for $\gamma = 0$ from (8.81) is therefore identical to the F-statistic for $c = 0$ from the GNR given in (8.82). Hence, "Tests based on the GNR are equivalent to variable addition tests when the latter are applicable," see Davidson and MacKinnon (1993, p. 194).

Note also, that the nR_u^2 test statistic for H_0; $\gamma = 0$ based on the GNR in (8.82) is exactly the LM statistic based on running the restricted least squares residuals of y on X on the unrestricted set of regressors X and Z in (8.81). If X has a constant, then the uncentered R^2 is equal to the centered R^2 printed by the regression.

Computational Warning: It is tempting to base tests on the OLS residuals $\widehat{u} = \bar{P}_X y$ by simply regressing them on the test regressors Z. This is equivalent to running the GNR *without* the X variables on the right hand side of (8.82) yielding test-statistics that are *too small*.

Functional Form

Davidson and MacKinnon (1993, p. 195) show that the RESET with $y_t = X_t\beta + \widehat{y}_t^2 c +$ residual which is based on testing for $c = 0$ is equivalent to testing for $\theta = 0$ using the nonlinear model $y_t = X_t\beta(1 + \theta X_t\beta) + u_t$. In this case, it is easy to verify from (8.74) that the GNR is

$$y_t - X_t\beta(1 + \theta X_t\beta) = (2\theta(X_t\beta)X_t + X_t)b + (X_t\beta)^2 c + \text{residual}$$

At $\theta = 0$ and $\beta = \widehat{\beta}_{OLS}$, the GNR becomes $(y_t - X_t\widehat{\beta}_{OLS}) = X_t b + (X_t\widehat{\beta}_{OLS})^2 c +$ residual. The t-statistic on $c = 0$ is equivalent to that from the RESET regression given in section 8.3, see problem 25.

Testing for Serial Correlation

Suppose that the null hypothesis is the nonlinear regression model given in (8.71), and the alternative is the model $y_t = x_t(\beta) + \nu_t$ with $\nu_t = \rho\nu_{t-1} + u_t$ where $u_t \sim \text{IID}(0, \sigma^2)$. Conditional on the first observation, the alternative model can be written as

$$y_t = x_t(\beta) + \rho(y_{t-1} - x_{t-1}(\beta)) + u_t$$

The GNR test for H_0; $\rho = 0$, computes the derivatives of this regression function with respect to β and ρ evaluated at the restricted estimates under the null hypothesis, i.e., $\rho = 0$ and $\beta = \widehat{\beta}_{NLS}$ (the nonlinear least squares estimate of β assuming no serial correlation). Those yield $X_t(\widehat{\beta}_{NLS})$ and $(y_{t-1} - x_{t-1}(\widehat{\beta}_{NLS}))$ respectively. Therefore, the GNR runs $\widehat{u}_t = y_t - x_t(\widehat{\beta}_{NLS}) = X_t(\widehat{\beta}_{NLS})b + c\widehat{u}_{t-1} +$ residual, and tests that $c = 0$. If the regression model is linear, this reduces to running ordinary least squares residuals on their lagged values in addition to the regressors in the model. This is exactly the Breusch and Godfrey test for first-order serial correlation considered in Chapter 5. For other applications as well as benefits and limitations of the GNR, see Davidson and MacKinnon (1993).

8.5 Testing Linear versus Log-Linear Functional Form[5]

In many economic applications where the explanatory variables take only positive values, econometricians must decide whether a linear or log-linear regression model is appropriate. In general, the linear model is given by

$$y_i = \sum_{j=1}^{k} \beta_j X_{ij} + \sum_{s=1}^{\ell} \gamma_s Z_{is} + u_i \quad i = 1, 2, \ldots, n \tag{8.84}$$

and the log-linear model is

$$\log y_i = \sum_{j=1}^{k} \beta_j \log X_{ij} + \sum_{s=1}^{\ell} \gamma_s Z_{is} + u_i \quad i = 1, 2, \ldots, n \tag{8.85}$$

with $u_i \sim \text{NID}(0, \sigma^2)$. Note that, the log-linear model is general in that only the dependent variable y and a subset of the regressors, i.e., the X variables are subject to the logarithmic transformation. Of course, one could estimate both models and compare their log-likelihood values. This would tell us which model fits best, but not whether either is a valid specification.

Box and Cox (1964) suggested the following transformation

$$B(y_i, \lambda) = \begin{cases} \frac{y_i^{\lambda} - 1}{\lambda} & \text{when} \quad \lambda \neq 0 \\ \log y_i & \text{when} \quad \lambda = 0 \end{cases} \tag{8.86}$$

where $y_i > 0$. Note that for $\lambda = 1$, as long as there is constant in the regression, subjecting the linear model to a Box-Cox transformation is equivalent to not transformation yields the log-linear regression. Therefore, the following Box-Cox model regression. Therefore, the following Box-Cox model

$$B(y_i, \lambda) = \sum_{j=1}^{k} \beta_j B(X_{ij}, \lambda) + \sum_{s=1}^{\ell} \gamma_s Z_{is} + u_i \tag{8.87}$$

encompasses as special cases the linear and log-linear models given in (8.84) and (8.85), respectively. Box and Cox (1964) suggested estimating these models by ML and using the LR test to test (8.84) and (8.85) against (8.87). However, estimation of (8.87) is computationally burdensome, see Davidson and MacKinnon (1993). Instead, we give an LM test involving a Double Length Regression (DLR) due to Davidson and MacKinnon (1985) that is easier to compute. In fact, Davidson and MacKinnon (1993, p. 510) point out that "everything that one can do with the Gauss-Newton Regression for nonlinear regression models can be done with the DLR for models involving transformations of the dependent variable." The GNR is not applicable in cases where the dependent variable is subjected to a nonlinear transformation, so one should use a DLR in these cases. Conversely, in cases where the GNR is valid, there is no need to run the DLR, since in these cases the latter is equivalent to the GNR.

For the linear model (8.84), the null hypothesis is that $\lambda = 1$. In this case, Davidson and MacKinnon suggest running a regression with 2n observations where the dependent variable has observations $(e_1/\widehat{\sigma}, \ldots, e_n/\widehat{\sigma}, 1, \ldots, 1)'$, i.e., the first n observations are the OLS residuals from (8.84) divided by the MLE of σ, where $\widehat{\sigma}^2_{mle} = e'e/n$. The second n observations are all equal to 1. The $2n$ observations for the regressors have typical elements:

$$\begin{array}{llll}
\text{for } \beta_j\text{: } X_{ij} - 1 & \text{for } i = 1, \ldots, n & \text{and} \quad 0 & \text{for the second } n \text{ elements} \\
\text{for } \gamma_s\text{: } Z_{is} & \text{for } i = 1, \ldots, n & \text{and} \quad 0 & \text{for the second } n \text{ elements} \\
\text{for } \sigma\text{: } e_i/\widehat{\sigma} & \text{for } i = 1, \ldots, n & \text{and} \quad -1 & \text{for the second } n \text{ elements} \\
\text{for } \lambda\text{: } \sum_{j=1}^{k} \widehat{\beta}_j (X_{ij} \log X_{ij} - X_{ij} + 1) - (y_i \log\, y_i - y_{i+1}) & & & \text{for } i = 1, \ldots, n \\
 & & \text{and} \quad \widehat{\sigma} \log y_i & \text{for the second } n \text{ elements}
\end{array}$$

The explained sum of squares for this DLR provides an asymptotically valid test for $\lambda = 1$. This will be distributed as χ^2_1 under the null hypothesis.

Similarly, when testing the log-linear model (8.85), the null hypothesis is that $\lambda = 0$. In this case, the dependent variable of the DLR has observations $(\widetilde{e}_1/\widetilde{\sigma}, \widetilde{e}_2/\widetilde{\sigma}, \ldots, \widetilde{e}_n/\widetilde{\sigma}, 1, \ldots, 1)'$, i.e., the first n observations are the OLS residuals from (8.85) divided by the MLE for σ, i.e., $\widetilde{\sigma}$ where $\widetilde{\sigma}^2 = \widetilde{e}'\widetilde{e}/n$. The second n observations are all equal to 1. The $2n$ observations for the regressors have typical elements:

$$\begin{array}{llll}
\text{for } \beta_j\text{: } \log X_{ij} & \text{for } i = 1, \ldots, n & \text{and} \quad 0 & \text{for the second } n \text{ elements} \\
\text{for } \gamma_s\text{: } Z_{is} & \text{for } i = 1, \ldots, n & \text{and} \quad 0 & \text{for the second } n \text{ elements} \\
\text{for } \sigma\text{: } \widetilde{e}_i/\widetilde{\sigma} & \text{for } i = 1, \ldots, n & \text{and} \quad -1 & \text{for the second } n \text{ elements} \\
\text{for } \lambda\text{: } \frac{1}{2}\sum_{j=1}^{k} \widetilde{\beta}_j (\log X_{ij})^2 - \frac{1}{2}(\log y_i)^2 & \text{for } i = 1, \ldots, n & & \\
 & \text{and} \quad \widetilde{\sigma} \log y_i & & \text{for the second } n \text{ elements}
\end{array}$$

The explained sum of squares from this DLR provides an asymptotically valid test for $\lambda = 0$. This will be distributed as χ^2_1 under the null hypothesis.

For the cigarette data given in Table 3.2, the linear model is given by $C = \beta_0 + \beta_1 P + \beta_2 Y + u$ whereas the log-linear model is given by $\log C = \gamma_0 + \gamma_1 \log P + \gamma_2 \log Y + \epsilon$ and the Box-Cox model is given by $B(C, \lambda) = \delta_0 + \delta_1 B(P, \lambda) + \delta_2 B(Y, \lambda) + \nu$, where $B(C, \lambda)$ is defined in (8.86). In this case, the DLR which tests the hypothesis that H_0; $\lambda = 1$, i.e., the model is linear, gives an explained sum of squares equal to 15.55. This is greater than a $\chi^2_{1,0.05} = 3.84$ and is therefore significant at the 5% level. Similarly the DLR that tests the hypothesis that H_0; $\lambda = 0$, i.e., the model is log-linear, gives an explained sum of squares equal to 8.86. This is also greater than $\chi^2_{1,0.05} = 3.84$ and is therefore significant at the 5% level. In this case, both the linear and log-linear models are rejected by the data.

Finally, it is important to note that there are numerous other tests for testing linear and log-linear models and the interested reader should refer to Davidson and MacKinnon (1993).

Notes

1. This section is based on Belsley, Kuh and Welsch (1980).
2. Other residuals that are linear unbiased with a scalar covariance matrix (LUS) are the BLUS residuals suggested by Theil (1971). Since we are explicitly dealing with time-series data, we use subscript t rather than i to index observations and T rather than n to denote the sample size.
3. Ramsey's (1969) initial formulation was based on BLUS residuals, but Ramsey and Schmidt (1976) showed that this is equivalent to using OLS residuals.
4. This section is based on Davidson and MacKinnon (1993, 2001).
5. This section is based on Davidson and MacKinnon (1993, pp. 502-510).

Problems

1. We know that $H = P_X$ is idempotent. Also, $(I_n - P_X)$ is idempotent. Therefore, $b'Hb \geq 0$ for any arbitrary vector b. Using these facts, show for $b' = (1, 0, \ldots, 0)$ that $0 \leq h_{11} \leq 1$. Deduce that $0 \leq h_{ii} \leq 1$ for $i = 1, \ldots, n$.
2. For the *simple regression with no constant* $y_i = x_i\beta + u_i$ for $i = 1, \ldots, n$
   (a) What is h_{ii}? Verify that $\sum_{i=1}^{n} h_{ii} = 1$.
   (b) What is $\widehat{\beta} - \widehat{\beta}_{(i)}$, see (8.13)? What is $s^2_{(i)}$ in terms of s^2 and e_i^2, see (8.18)? What is $DFBETAS_{ij}$, see (8.19)?
   (c) What are $DFFIT_i$ and $DFFITS_i$, see (8.21) and (8.22)?
   (d) What is Cook's distance measure $D_i^2(s)$ for this simple regression with no intercept, see (8.24)?
   (e) Verify that (8.27) holds for this simple regression with no intercept. What is $COVRATIO_i$, see (8.26)?
3. From the definition of $s^2_{(i)}$ in (8.17), substitute (8.13) in (8.17) and verify (8.18).
4. Consider the augmented regression given in (8.5) $y = X\beta^* + d_i\varphi + u$ where φ is a scalar and $d_i = 1$ for the i-th observation and 0 otherwise. Using the Frisch-Waugh Lovell Theorem given in section 7.3, verify that
   (a) $\widehat{\beta}^* = (X'_{(i)}X_{(i)})^{-1}X'_{(i)}y_{(i)} = \widehat{\beta}_{(i)}$.
   (b) $\widehat{\varphi} = (d_i'\bar{P}_X d_i)^{-1}d_i'\bar{P}_X y = e_i/(1 - h_{ii})$ where $\bar{P}_X = I - P_X$.
   (c) Residual Sum of Squares from (8.5) = (Residual Sum of Squares with d_i deleted) $- e_i^2/(1 - h_{ii})$.
   (d) Assuming Normality of u, show that the t-statistic for testing $\varphi = 0$ is $t = \widehat{\varphi}/s.e.(\widehat{\varphi}) = e_i^*$ as given in (8.3).

5. Consider the augmented regression $y = X\beta^* + \bar{P}_X D_p \varphi^* + u$, where D_p is an $n \times p$ matrix of dummy variables for the p suspected observations. Note that $\bar{P}_X D_p$ rather than D_p appear in this equation. Compare with (8.6). Let $e_p = D_p' e$, then $E(e_p) = 0$, $\text{var}(e_p) = \sigma^2 D_p' \bar{P}_X D_p$. Verify that

   (a) $\widehat{\beta}^* = (X'X)^{-1}X'y = \widehat{\beta}_{OLS}$ and

   (b) $\widehat{\varphi}^* = (D_p'\bar{P}_X D_p)^{-1} D_p' \bar{P}_X y = (D_p'\bar{P}_X D_p)^{-1} D_p' e = (D_p'\bar{P}_X D_p)^{-1} e_p$.

   (c) Residual Sum of Squares = (Residual Sum of Squares with D_p deleted) $- e_p'(D_p'\bar{P}_X)D_p^{-1}e_p$. Using the Frisch-Waugh Lovell Theorem show this residual sum of squares is the same as that for (8.6).

   (d) Assuming normality of u, verify (8.7) and (8.9).

   (e) Repeat this exercise for problem 4 with $\bar{P}_X d_i$ replacing d_i. What do you conclude?

6. Using the updating formula in (8.11), verify (8.12) and deduce (8.13).

7. Verify that Cook's distance measure given in (8.25) is related to $DFFITS_i(\sigma)$ as follows: $DFFITS_i(\sigma) = \sqrt{k} D_i(\sigma)$.

8. Using the matrix identity $\det(I_k - ab') = 1 - b'a$, where a and b are column vectors of dimension k, prove (8.27). **Hint:** Use $a = x_i$ and $b' = x_i'(X'X)^{-1}$ and the fact that $\det[X_{(i)}'X_{(i)}] = \det[\{I_k - x_i x_i'(X'X)^{-1}\}X'X]$.

9. For the cigarette data given in Table 3.2

   (a) Replicate the results in Table 8.2.

   (b) For the New Hampshire observation (NH), compute $\widetilde{e}_{NH}$, e^*_{NH}, $\widehat{\beta} - \widehat{\beta}_{(NH)}$, $DFBETAS_{NH}$, $DFFIT_{NH}$, $DFFITS_{NH}$, $D^2_{NH}(s)$, $COVRATIO_{NH}$, and $FVARATIO_{NH}$.

   (c) Repeat the calculations in part (b) for the following states: AR, CT, NJ and UT.

   (d) What about the observations for NV, ME, NM and ND? Are they influential?

10. For the Consumption-Income data given in Table 5.1, compute

    (a) The internal studentized residuals $\widetilde{e}$ given in (8.1).

    (b) The externally studentized residuals e^* given in (8.3).

    (c) Cook's statistic given in (8.25).

    (d) The leverage of each observation h.

    (e) The *DFFITS* given in (8.22).

    (f) The *COVRATIO* given in (8.28).

    (g) Based on the results in parts (a) to (f), identify the observations that are influential.

11. Repeat problem 10 for the 1982 data on earnings used in Chapter 4. This data is provided on the Springer web site as EARN.ASC.

12. Repeat problem 10 for the Gasoline data provided on the Springer web site as GASOLINE.DAT. Use the gasoline demand model given in Chapter 10, section 5. Do this for Austria and Belgium separately.

13. *Independence of Recursive Residuals.*

    (a) Using the updating formula given in (8.11) with $A = (X_t'X_t)$ and $a = -b = x_{t+1}'$, verify (8.31).

(b) Using (8.31), verify (8.32).

(c) For $u_t \sim \text{IIN}(0, \sigma^2)$ and w_{t+1} defined in (8.30) verify (8.33). **Hint:** define $v_{t+1}=\sqrt{f_{t+1}}w_{t+1}$. From (8.30), we have

$$v_{t+1} = \sqrt{f_{t+1}}w_{t+1} = y_{t+1} - x'_{t+1}\widehat{\beta}_t = x'_{t+1}(\beta - \widehat{\beta}_t) + u_{t+1} \quad \text{for} \quad t = k, \ldots, T-1$$

Since f_{t+1} is fixed, it suffices to show that $\text{cov}(v_{t+1}, v_{s+1}) = 0$ for $t \neq s$.

14. *Recursive Residuals are Linear Unbiased With Scalar Covariance Matrix (LUS).*

(a) Verify that the $(T-k)$ recursive residuals defined in (8.30) can be written in vector form as $w = Cy$ where C is defined in (8.34). This shows that the recursive residuals are linear in y.

(b) Show that C satisfies the three properties given in (8.35) i.e., $CX = 0$, $CC' = I_{T-k}$, and $C'C = \bar{P}_X$. Prove that $CX = 0$ means that the recursive residuals are unbiased with zero mean. Prove that the $CC' = I_{T-k}$ means that the recursive residuals have a scalar covariance matrix. Prove that $C'C = \bar{P}_X$ means that the sum of squares of $(T-k)$ recursive residuals is equal to the sum of squares of T least squares residuals.

(c) If the true disturbances $u \sim N(0, \sigma^2 I_T)$, prove that the recursive residuals $w \sim N(0, \sigma^2 I_{T-k})$ using parts (a) and (b).

(d) Verify (8.36), i.e., show that $RSS_{t+1} = RSS_t + w^2_{t+1}$ for $t = k, \ldots, T-1$ where $RSS_t = (Y_t - X_t\widehat{\beta}_t)'(Y_t - X_t\widehat{\beta}_t)$.

15. *The Harvey and Collier (1977) Misspecification t-test as a Variable Additions Test.* This is based on Wu (1993).

(a) Show that the F-statistic for testing H_0; $\gamma = 0$ versus $\gamma \neq 0$ in (8.44) is given by

$$F = \frac{y'\bar{P}_X y - y'\bar{P}_{[X,z]}y}{y'\bar{P}_{[X,z]}y/(T-k-1)} = \frac{y'P_z y}{y'(\bar{P}_X - P_z)y/(T-k-1)}$$

and is distributed as $F(1, T-k-1)$ under the null hypothesis.

(b) Using the properties of C given in (8.35), show that the F-statistic given in part (a) is the square of the Harvey and Collier (1977) t-statistic given in (8.43).

16. For the Gasoline data for Austria given on the Springer web site as GASOLINE.DAT and the model given in Chapter 10, section 5, compute:

(a) The recursive residuals given in (8.30).

(b) The CUSUM given in (8.46) and plot it against r.

(c) Draw the 5% upper and lower lines given below (8.46) and see whether the CUSUM crosses these boundaries.

(d) The post-sample predictive test for 1978. Verify that computing it from (8.38) or (8.40) yields the same answer.

(e) The modified von Neuman ratio given in (8.42).

(f) The Harvey and Collier (1977) functional misspecification test given in (8.43).

17. *The Differencing Test in a Regression with Equicorrelated Disturbances.* This is based on Baltagi (1990). Consider the time-series regression

$$Y = \iota_T \alpha + X\beta + u \tag{1}$$

where ι_T is a vector of ones of dimension T. X is $T \times K$ and $[\iota_T, X]$ is of full column rank. $u \sim (0, \Omega)$ where Ω is positive definite. Differencing this model, we get

$$DY = DX\beta + Du \tag{2}$$

where D is a $(T-1) \times T$ matrix given below (8.50). Maeshiro and Wichers (1989) show that GLS on (1) yields through partitioned inverse:

$$\widehat{\beta} = (X'LX)^{-1}X'LY \tag{3}$$

where $L = \Omega^{-1} - \Omega^{-1}\iota_T(\iota_T'\Omega^{-1}\iota_T)^{-1}\iota_T'\Omega^{-1}$. Also, GLS on (2) yields

$$\widetilde{\beta} = (X'MX)^{-1}X'MY \tag{4}$$

where $M = D'(D\Omega D')^{-1}D$. Finally, they show that $M = L$, and GLS on (2) is equivalent to GLS on (1) as long as there is an intercept in (1).

Consider the special case of equicorrelated disturbances

$$\Omega = \sigma^2[(1-\rho)I_T + \rho J_T] \tag{5}$$

where I_T is an identity matrix of dimension T and J_T is a matrix of ones of dimension T.

(a) Derive the L and M matrices for the equicorrelated case, and verify the Maeshiro and Wichers result for this special case.

(b) Show that for the equicorrelated case, the differencing test given by Plosser, Schwert, and White (1982) can be obtained as the difference between the OLS and GLS estimators of the differenced equation (2). **Hint:** See the solution by Koning (1992).

18. For the 1982 data on earnings used in Chapter 4, provided as EARN.ASC on the Springer web site compute Ramsey's (1969) RESET and the Thursby and Schmidt (1977) variant of this test.

19. Repeat problem 18 for the Hedonic housing data given on the Springer web site as HEDONIC.XLS.

20. Repeat problem 18 for the cigarette data given in Table 3.2.

21. Repeat problem 18 for the Gasoline data for Austria given on the Springer web site as GASOLINE.DAT. Use the model given in Chapter 10, section 5. Also compute the PSW differencing test given in (8.54).

22. Use the 1982 data on earnings used in Chapter 4, and provided on the Springer web site as EARN.ASC. Consider the two competing non-nested models

$$\begin{aligned} H_0;\ \log(\text{wage}) &= \beta_0 + \beta_1 ED + \beta_2 EXP + \beta_3 EXP^2 + \beta_4 WKS \\ &\quad + \beta_5 MS + \beta_6 FEM + \beta_7 BLK + \beta_8 UNION + u \end{aligned}$$

$$\begin{aligned} H_1;\ \log(\text{wage}) &= \gamma_0 + \gamma_1 ED + \gamma_2 EXP + \gamma_3 EXP^2 + \gamma_4 WKS \\ &\quad + \gamma_5 OCC + \gamma_6 SOUTH + \gamma_7 SMSA + \gamma_8 IND + \epsilon \end{aligned}$$

Compute:

(a) The Davidson and MacKinnon (1981) J-test for H_0 versus H_1.

(b) The Fisher and McAleer (1981) JA-test for H_0 versus H_1.

(c) Reverse the roles of H_0 and H_1 and repeat parts (a) and (b).

(d) Both H_0 and H_1 can be artificially nested in the model used in Chapter 4. Using the F-test given in (8.62), test for H_0 versus this augmented model. Repeat for H_1 versus this augmented model. What do you conclude?

23. For the Consumption-Income data given in Table 5.1,

(a) Test the hypothesis that the Consumption model is linear Box-Cox alternative.

(b) Test the hypothesis that the Consumption model is log-linear general Box-Cox alternative.

24. Repeat problem 23 for the Cigarette data given in Table 3.2.

25. *RESET as a Gauss-Newton Regression.* This is based on Baltagi (1998). Davidson and MacKinnon (1993) showed that Ramsey's (1969) regression error specification test (RESET) can be derived as a Gauss-Newton Regression. This problem is a simple extension of their results. Suppose that the linear regression model under test is given by:

$$y_t = X_t'\beta + u_t \qquad t = 1, 2, \ldots, T \tag{1}$$

where β is a $k \times 1$ vector of unknown parameters. Suppose that the alternative is the nonlinear regression model between y_t and X_t:

$$y_t = X_t'\beta[1 + \theta(X_t'\beta) + \gamma(X_t'\beta)^2 + \lambda(X_t'\beta)^3] + u_t, \tag{2}$$

where θ, γ, and λ are unknown scalar parameters. It is well known that Ramsey's (1969) RESET is obtained by regressing y_t on X_t, $\widehat{y}_t^2$, $\widehat{y}_t^3$ and $\widehat{y}_t^4$ and by testing that the coefficients of all powers of $\widehat{y}_t$ are jointly zero. Show that this RESET can be derived from a Gauss-Newton Regression on (2), which tests $\theta = \gamma = \lambda = 0$.

References

This chapter is based on Belsley, Kuh and Welsch (1980), Johnston (1984), Maddala (1992) and Davidson and MacKinnon (1993). Additional references are the following:

Baltagi, B.H. (1990), "The Differencing Test in a Regression with Equicorrelated Disturbances," *Econometric Theory*, Problem 90.4.5, 6: 488.

Baltagi, B.H. (1998), "Regression Specification Error Test as A Gauss-Newton Regression," *Econometric Theory*, Problem 98.4.3, 14: 526.

Belsley, D.A., E. Kuh and R.E. Welsch (1980), *Regression Diagnostics* (Wiley: New York).

Box, G.E.P. and D.R. Cox (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Series B, 26: 211-252.

Brown, R.L., J. Durbin, and J.M. Evans (1975), "Techniques for Testing the Constancy of Regression Relationships Over Time," *Journal of the Royal Statistical Society* 37:149-192.

Chesher, A. and R. Spady (1991), "Asymptotic Expansions of the Information Matrix Test Statistic," *Econometrica* 59: 787-815.

Cook, R.D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics* 19:15-18.

Cook, R.D. and S. Weisberg (1982), *Residuals and Influences in Regression* (Chapman and Hall: New York).

Cox, D.R. (1961), "Tests of Separate Families of Hypotheses," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1: 105-123.

Davidson, R., L.G. Godfrey and J.G. MacKinnon (1985), "A Simplified Version of the Differencing Test," *International Economic Review*, 26: 639-47.

Davidson, R. and J.G. MacKinnon (1981), "Several Tests for Model Specification in the Presence of Alternative Hypotheses," *Econometrica*, 49: 781-793.

Davidson, R. and J.G. MacKinnon (1985), "Testing Linear and Loglinear Regressions Against Box-Cox Alternatives," *Canadian Journal of Economics*, 18: 499-517.

Davidson, R. and J.G. MacKinnon (1992), "A New Form of the Information Matrix Test," *Econometrica*, 60: 145-157.

Davidson, R. and J.G. MacKinnon (2001), "Artificial Regressions," Chapter 1 in Baltagi, B.H. (ed.) *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Fisher, G.R. and M. McAleer (1981), "Alternative Procedures and Associated Tests of Significance for Non-Nested Hypotheses," *Journal of Econometrics*, 16: 103-119.

Gentleman, J.F. and M.B. Wilk (1975), "Detecting Outliers II: Supplementing the Direct Analysis of Residuals," *Biometrics*, 31: 387-410.

Godfrey, L.G. (1988), *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches* (Cambridge University Press: Cambridge).

Hall, A. (1987), "The Information Matrix Test for the Linear Model," *Review of Economic Studies*, 54: 257-263.

Harvey, A.C. (1976), "An Alternative Proof and Generalization of a Test for Structural Change," *The American Statistician*, 30: 122-123.

Harvey, A.C. (1990), *The Econometric Analysis of Time Series* (MIT Press: Cambridge).

Harvey, A.C. and P. Collier (1977), "Testing for Functional Misspecification in Regression Analysis," *Journal of Econometrics*, 6: 103-119.

Harvey, A.C. and G.D.A. Phillips (1974), "A Comparison of the Power of Some Tests for Heteroskedasticity in the General Linear Model," *Journal of Econometrics*, 2: 307-316.

Hausman, J. (1978), "Specification Tests in Econometrics," *Econometrica*, 46: 1251-1271.

Koning, R.H. (1992), "The Differencing Test in a Regression with Equicorrelated Disturbances," *Econometric Theory*, Solution 90.4.5, 8: 155-156.

Krämer, W. and H. Sonnberger (1986), *The Linear Regression Model Under Test* (Physica-Verlag: Heidelberg).

Krasker, W.S., E. Kuh and R.E. Welsch (1983), "Estimation for Dirty Data and Flawed Models," Chapter 11 in *Handbook of Econometrics*, Vol. I, eds. Z. Griliches and M.D. Intrilligator, Amsterdam, North-Holland.

Maeshiro, A. and R. Wichers (1989), "On the Relationship Between the Estimates of Level Models and Difference Models," *American Journal of Agricultural Economics*, 71: 432-434.

Orme, C. (1990), "The Small Sample Performance of the Information Matrix Test," *Journal of Econometrics*, 46: 309-331.

Pagan, A.R. and A.D. Hall (1983), "Diagnostic Tests as Residual Analysis," *Econometric Reviews*, 2: 159-254.

Pesaran, M.H. and M. Weeks (2001), "Nonnested Hypothesis Testing: A Overview," Chapter 13 in Baltagi, B.H. (ed.) *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Phillips, G.D.A. and A.C. Harvey (1974), "A Simple Test for Serial Correlation in Regression Analysis," *Journal of the American Statistical Association*, 69: 935-939.

Plosser, C.I., G.W. Schwert, and H. White (1982), "Differencing as a Test of Specification," *International Economic Review*, 23: 535-552.

Ramsey, J.B. (1969), "Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis," *Journal of the Royal Statistics Society*, Series B, 31: 350-371.

Ramsey, J.B. and P. Schmidt (1976), "Some Further Results in the Use of OLS and BLUS Residuals in Error Specification Tests," *Journal of the American Statistical Association*, 71: 389-390.

Schmidt, P. (1976), *Econometrics* (Marcel Dekker: New York).

Theil, H. (1971), *Principles of Econometrics* (Wiley: New York).

Thursby, J. and P. Schmidt (1977), "Some Properties of Tests for Specification Error in a Linear Regression Model," *Journal of the American Statistical Association*, 72: 635-641.

Utts, J.M. (1982), "The Rainbow Test for Lack of Fit in Regression," *Communications in Statistics*, 11: 2801-2815.

Velleman, P. and R. Welsch (1981), "Efficient Computing of Regression Diagnostics," *The American Statistician*, 35: 234-242.

White, H. (1980), "A Heteroskedasticity–Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48: 817-838.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50: 1-25.

Wooldridge, J.M. (1995), "Diagnostic Testing," Chapter 9 in B.H. Baltagi (ed.) *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Wu, P. (1993), "Variable Addition Test," *Econometric Theory*, Problem 93.1.2, 9: 145-146.

CHAPTER 9
Generalized Least Squares

9.1 Introduction

This chapter considers a more general variance covariance matrix for the disturbances. In other words, $u \sim (0, \sigma^2 I_n)$ is relaxed so that $u \sim (0, \sigma^2 \Omega)$ where Ω is a positive definite matrix of dimension $(n \times n)$. First Ω is assumed known and the BLUE for β is derived. This estimator turns out to be different from $\widehat{\beta}_{OLS}$, and is denoted by $\widehat{\beta}_{GLS}$, the Generalized Least Squares estimator of β. Next, we study the properties of $\widehat{\beta}_{OLS}$ under this nonspherical form of the disturbances. It turns out that the OLS estimates are still unbiased and consistent, but their standard errors as computed by standard regression packages are biased and inconsistent and lead to misleading inference. Section 9.3 studies some special forms of Ω and derive the corresponding BLUE for β. It turns out that heteroskedasticity and serial correlation studied in Chapter 5 are special cases of Ω. Section 9.4 introduces normality and derives the maximum likelihood estimator. Sections 9.5 and 9.6 study the way in which test of hypotheses and prediction get affected by this general variance-covariance assumption on the disturbances. Section 9.7 studies the properties of this BLUE for β when Ω is unknown, and is replaced by a consistent estimator. Section 9.8 studies what happens to the W, LR and LM statistics when $u \sim N(0, \sigma^2 \Omega)$. Section 9.9 gives another application of GLS to spatial autocorrelation.

9.2 Generalized Least Squares

The regression equation did not change, only the variance-covariance matrix of the disturbances. It is now $\sigma^2 \Omega$ rather than $\sigma^2 I_n$. However, we can rely once again on a result from matrix algebra to transform our nonspherical disturbances back to spherical form, see the Appendix to Chapter 7. This result states that for every positive definite matrix Ω, there exists a *nonsingular* matrix P such that $PP' = \Omega$. In order to use this result, we transform the original model

$$y = X\beta + u \tag{9.1}$$

by premultiplying it by P^{-1}. We get

$$P^{-1}y = P^{-1}X\beta + P^{-1}u \tag{9.2}$$

Defining y^* as $P^{-1}y$ and X^* and u^* similarly, we have

$$y^* = X^*\beta + u^* \tag{9.3}$$

with u^* having 0 mean and $\text{var}(u^*) = P^{-1}\text{var}(u)P^{-1\prime} = \sigma^2 P^{-1}\Omega P'^{-1} = \sigma^2 P^{-1}PP'P'^{-1} = \sigma^2 I_n$. Hence, the variance-covariance of the disturbances in (9.3) is a scalar times an identity matrix. Therefore, using the results of Chapter 7, the BLUE for β in (9.1) is OLS on the transformed model in (9.3)

$$\widehat{\beta}_{BLUE} = (X^{*\prime}X^*)^{-1}X^{*\prime}y^* = (X'P^{-1\prime}P^{-1}X)^{-1}X'P^{-1\prime}P^{-1}y = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \tag{9.4}$$

with $\text{var}(\widehat{\beta}_{BLUE}) = \sigma^2(X^{*\prime}X^*)^{-1} = \sigma^2(X'\Omega^{-1}X)^{-1}$. This $\widehat{\beta}_{BLUE}$ is known as $\widehat{\beta}_{GLS}$. Define $\Sigma = E(uu') = \sigma^2\Omega$, then Σ differs from Ω only by the positive scalar σ^2. One can easily verify that $\widehat{\beta}_{GLS}$ can be alternatively written as $\widehat{\beta}_{GLS} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$ and that $\text{var}(\widehat{\beta}_{GLS}) = (X'\Sigma^{-1}X)^{-1}$. Just substitute $\Sigma^{-1} = \Omega^{-1}/\sigma^2$ in this last expression for $\widehat{\beta}_{GLS}$ and verify that this yields (9.4).

It is clear that $\widehat{\beta}_{GLS}$ differs from $\widehat{\beta}_{OLS}$. In fact, since $\widehat{\beta}_{OLS}$ is still a linear unbiased estimator of β, the Gauss-Markov Theorem states that it must have a variance larger than that of $\widehat{\beta}_{GLS}$. Using equation (7.5) from Chapter 7, i.e., $\widehat{\beta}_{OLS} = \beta + (X'X)^{-1}X'u$ it is easy to show that

$$\text{var}(\widehat{\beta}_{OLS}) = \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1} \tag{9.5}$$

Problem 1 shows that $\text{var}(\widehat{\beta}_{OLS}) - \text{var}(\widehat{\beta}_{GLS})$ is a positive semi-definite matrix. Note that $\text{var}(\widehat{\beta}_{OLS})$ is no longer $\sigma^2(X'X)^{-1}$, and hence a regression package that is programmed to compute $s^2(X'X)^{-1}$ as an estimate of the variance of $\widehat{\beta}_{OLS}$ is using the wrong formula. Furthermore, problem 2 shows that $E(s^2)$ is not in general σ^2. Hence, the regression package is also wrongly estimating σ^2 by s^2. Two wrongs do not make a right, and the estimate of $\text{var}(\widehat{\beta}_{OLS})$ is biased. The direction of this bias depends upon the form of Ω and the X matrix. (We saw some examples of this bias under heteroskedasticity and serial correlation in Chapter 5). Hence, the standard errors and t-statistics computed using this OLS regression are biased. Under heteroskedasticity, one can use the White (1980) robust standard errors for OLS. In this case, $\Sigma = \sigma^2\Omega$ in (9.5) is estimated by $\widehat{\Sigma} = diag[e_i^2]$ where e_i denotes the least squares residuals. The resulting t-statistics are robust to heteroskedasticity. Similarly Wald type statistics for H_0; $R\beta = r$ can be obtained based on $\widehat{\beta}_{OLS}$ by replacing $\sigma^2(X'X)^{-1}$ in (7.41) by (9.5) with $\widehat{\Sigma} = diag[e_i^2]$. In the presence of both serial correlation and heteroskedasticity, one can use the consistent covariance matrix estimate suggested by Newey and West (1987). This was discussed in Chapter 5.

To summarize, $\widehat{\beta}_{OLS}$ is no longer BLUE whenever $\Omega \neq I_n$. However, it is still unbiased and consistent. The last two properties do not rely upon the form of the variance-covariance matrix of the disturbances but rather on $E(u/X) = 0$ and plim $X'u/n = 0$. The standard errors of $\widehat{\beta}_{OLS}$ as computed by the regression package are biased and any test of hypothesis based on this OLS regression may be misleading.

So far we have not derived an estimator for σ^2. We know however, from the results in Chapter 7, that the transformed regression (9.3) yields a mean squared error that is an unbiased estimator for σ^2. Denote this by s^{*2} which is equal to the transformed OLS residual sum of squares divided by $(n-K)$. Let e^* denote the vector of OLS residuals from (9.3), this means that $e^* = y^* - X^*\widehat{\beta}_{GLS} = P^{-1}(y - X\widehat{\beta}_{GLS}) = P^{-1}e_{GLS}$ and

$$\begin{aligned} s^{*2} &= e^{*\prime}e^*/(n-K) = (y - X\widehat{\beta}_{GLS})'\Omega^{-1}(y - X\widehat{\beta}_{GLS})/(n-K) \\ &= e'_{GLS}\Omega^{-1}e_{GLS}/(n-K) \end{aligned} \tag{9.6}$$

Note that s^{*2} now depends upon Ω^{-1}.

Necessary and Sufficient Conditions for OLS to be Equivalent to GLS

There are several necessary and sufficient conditions for OLS to be equivalent to GLS, see Puntanen and Styan (1989) for a historical survey. For pedagogical reasons, we focus on the derivation of Milliken and Albohali (1984). Note that $y = P_X y + \bar{P}_X y$. Therefore, replacing y

in $\widehat{\beta}_{GLS}$ by this expression we get

$$\widehat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}[P_X y + \bar{P}_X y] = \widehat{\beta}_{OLS} + (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\bar{P}_X y$$

The last term is zero for every y if and only if

$$X'\Omega^{-1}\bar{P}_X = 0 \tag{9.7}$$

Therefore, $\widehat{\beta}_{GLS} = \widehat{\beta}_{OLS}$ if and only if (9.7) is true.

Another easy necessary and sufficient condition to check in practice is the following:

$$P_X\Omega = \Omega P_X \tag{9.8}$$

see Zyskind (1967). This involves Ω rather than Ω^{-1}. There are several applications in economics where these conditions are satisfied and can be easily verified, see Balestra (1970) and Baltagi (1989). We will apply these conditions in Chapter 10 on Seemingly Unrelated Regressions, Chapter 11 on simultaneous equations and Chapter 12 on panel data. See also problem 9.

9.3 Special Forms of Ω

If the disturbances are heteroskedastic but not serially correlated, then $\Omega = \text{diag}[\sigma_i^2]$. In this case, $P = \text{diag}[\sigma_i]$, $P^{-1} = \Omega^{-1/2} = \text{diag}[1/\sigma_i]$ and $\Omega^{-1} = \text{diag}[1/\sigma_i^2]$. Premultiplying the regression equation by $\Omega^{-1/2}$ is equivalent to dividing the i-th observation of this model by σ_i. This makes the new disturbance u_i/σ_i have 0 mean and homoskedastic variance σ^2, leaving properties like no serial correlation intact. The new regression runs $y_i^* = y_i/\sigma_i$ on $X_{ik}^* = X_{ik}/\sigma_i$ for $i = 1, 2, \ldots, n$, and $k = 1, 2, \ldots, K$. Specific assumptions on the form of these σ_i's were studied in the heteroskedasticity chapter.

If the disturbances follow an AR(1) process $u_t = \rho u_{t-1} + \epsilon_t$ for $t = 1, 2, \ldots, T$; with $|\rho| \leq 1$ and $\epsilon_t \sim$IID$(0, \sigma_\epsilon^2)$, then $\text{cov}(u_t, u_{t-s}) = \rho^s \sigma_u^2$ with $\sigma_u^2 = \sigma_\epsilon^2/(1-\rho^2)$. This means that

$$\Omega = \begin{bmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{T-1} \\ \rho & 1 & \rho & \ldots & \rho^{T-2} \\ : & : & : & & : \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \ldots & 1 \end{bmatrix} \tag{9.9}$$

and

$$\Omega^{-1} = \left(\frac{1}{1-\rho^2}\right) \begin{bmatrix} 1 & -\rho & 0 & \ldots & 0 & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \ldots & 0 & 0 & 0 \\ : & : & : & & : & : & : \\ 0 & 0 & 0 & \ldots & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \ldots & 0 & -\rho & 1 \end{bmatrix} \tag{9.10}$$

Then

$$P^{-1} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \ldots & 0 & 0 & 0 \\ -\rho & 1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & -\rho & 1 & \ldots & 0 & 0 & 0 \\ : & : & : & & : & : & : \\ 0 & 0 & 0 & \ldots & -\rho & 1 & 0 \\ 0 & 0 & 0 & \ldots & 0 & -\rho & 1 \end{bmatrix} \tag{9.11}$$

is the matrix that satisfies the following condition $P^{-1\prime}P^{-1} = (1-\rho^2)\Omega^{-1}$. Premultiplying the regression model by P^{-1} is equivalent to performing the Prais-Winsten transformation. In particular the first observation on y becomes $y_1^* = \sqrt{1-\rho^2}y_1$ and the remaining observations are given by $y_t^* = (y_t - \rho y_{t-1})$ for $t = 2, 3, \ldots, T$, with similar terms for the X's and the disturbances. Problem 3 shows that the variance covariance matrix of the transformed disturbances $u^* = P^{-1}u$ is $\sigma_\epsilon^2 I_T$.

Other examples where an explicit form for P^{-1} has been derived include, (i) the MA(1) model, see Balestra (1980); (ii) the AR(2) model, see Lempers and Kloek (1973); (iii) the specialized AR(4) model for quarterly data, see Thomas and Wallis (1971); and (iv) the error components model, see Fuller and Battese (1974) and Chapter 12.

9.4 Maximum Likelihood Estimation

Assuming that $u \sim N(0, \sigma^2\Omega)$, the new likelihood function can be derived keeping in mind that $u^* = P^{-1}u = \Omega^{-1/2}u$ and $u^* \sim N(0, \sigma^2 I_n)$. In this case

$$f(u_1^*, \ldots, u_n^*; \beta, \sigma^2) = (1/2\pi\sigma^2)^{n/2} \exp\{-u^{*\prime}u^*/2\sigma^2\} \tag{9.12}$$

Making the transformation $u = Pu^* = \Omega^{1/2}u^*$, we get

$$f(u_1, \ldots, u_n; \beta, \sigma^2) = (1/2\pi\sigma^2)^{n/2}|\Omega^{-1/2}| \exp\{-u'\Omega^{-1}u/2\sigma^2\} \tag{9.13}$$

where $|\Omega^{-1/2}|$ is the Jacobian of the inverse transformation. Finally, substituting $y = X\beta + u$ in (9.13), one gets the likelihood function

$$L(\beta, \sigma^2; \Omega) = (1/2\pi\sigma^2)^{n/2}|\Omega^{-1/2}| \exp\{-(y - X\beta)'\Omega^{-1}(y - X\beta)/2\sigma^2\} \tag{9.14}$$

since the Jacobian of this last transformation is 1. Knowing Ω, maximizing (9.14) with respect to β is equivalent to minimizing $u^{*\prime}u^*$ with respect to β. This means that $\widehat{\beta}_{MLE}$ is the OLS estimate on the transformed model, i.e., $\widehat{\beta}_{GLS}$. From (9.14), we see that this RSS is a weighted one with the weight being the inverse of the variance covariance matrix of the disturbances. Similarly, maximizing (9.14) with respect to σ^2 gets $\widehat{\sigma}^2_{MLE}$ = the OLS residual sum of squares of the transformed regression (9.3) divided by n. From (9.6) this can be written as $\widehat{\sigma}^2_{MLE} = e^{*\prime}e^*/n = (n-K)s^{*2}/n$. The distributions of these maximum likelihood estimates can be derived from the transformed model using the results in Chapter 7. In fact, $\widehat{\beta}_{GLS} \sim N(\beta, \sigma^2(X'\Omega^{-1}X)^{-1})$ and $(n-K)s^{*2}/\sigma^2 \sim \chi^2_{n-K}$.

9.5 Test of Hypotheses

In order to test H_0; $R\beta = r$, under the general variance-covariance matrix assumption, one can revert to the transformed model (9.3) which has a scalar identity variance-covariance matrix and use the test statistic derived in Chapter 7

$$(R\widehat{\beta}_{GLS} - r)'[R(X^{*\prime}X^*)^{-1}R']^{-1}(R\widehat{\beta}_{GLS} - r)/\sigma^2 \sim \chi^2_g \tag{9.15}$$

Note that $\widehat{\beta}_{GLS}$ replaces $\widehat{\beta}_{OLS}$ and X^* replaces X. Replacing X^* by $P^{-1}X$, we get

$$(R\widehat{\beta}_{GLS} - r)'[R(X'\Omega^{-1}X)^{-1}R']^{-1}(R\widehat{\beta}_{GLS} - r)/\sigma^2 \sim \chi^2_g \tag{9.16}$$

This differs from its counterpart in the spherical disturbances model in two ways. $\widehat{\beta}_{GLS}$ replaces $\widehat{\beta}_{OLS}$, and $(X'\Omega^{-1}X)$ takes the place of $X'X$. One can also derive the restricted estimator based on the transformed model by simply replacing X^* by $P^{-1}X$ and the OLS estimator of β by its GLS counterpart. Problem 4 asks the reader to verify that the restricted GLS estimator is

$$\widehat{\beta}_{RGLS} = \widehat{\beta}_{GLS} - (X'\Omega^{-1}X)^{-1}R'[R(X'\Omega^{-1}X)^{-1}R']^{-1}(R\widehat{\beta}_{GLS} - r) \tag{9.17}$$

Furthermore, using the same analysis given in Chapter 7, one can show that (9.15) is in fact the Likelihood Ratio statistic and is equal to the Wald and Lagrangian Multiplier statistics, see Buse (1982). In order to operationalize these tests, we replace σ^2 by its unbiased estimate s^{*2}, and divide by g the number of restrictions. The resulting statistic is an $F(g, n-K)$ for the same reasons given in Chapter 7.

9.6 Prediction

How is prediction affected by nonspherical disturbances? Suppose we want to predict one period ahead. What has changed with a general Ω? For one thing, we now know that the period $(T+1)$ disturbance is correlated with the sample disturbances. Let us assume that this correlation is given by the $(T \times 1)$ vector $\omega = E(u_{T+1}u)$, problem 5 shows that the BLUP for y_{T+1} is

$$\widehat{y}_{T+1} = x'_{T+1}\widehat{\beta}_{GLS} + \omega'\Omega^{-1}(y - X\widehat{\beta}_{GLS})/\sigma^2 \tag{9.18}$$

The first term is as expected, however it is the second term that highlights the difference between the spherical and nonspherical model predictions. To illustrate this, let us look at the AR(1) case where $\text{cov}(u_t, u_{t-s}) = \rho^s\sigma_u^2$. This implies that $\omega' = \sigma_u^2(\rho^T, \rho^{T-1}, \ldots, \rho)$. Using Ω which is given in (9.9), one can show that ω is equal to $\rho\sigma_u^2$ multiplied by the last column of Ω. But $\Omega^{-1}\Omega = I_T$, therefore, Ω^{-1} times the last column of Ω gives the last column of the identity matrix, i.e., $(0, 0, \ldots, 1)'$. Substituting for the last column of Ω its expression $(\omega/\rho\sigma_u^2)$ one gets, $\Omega^{-1}(\omega/\rho\sigma_u^2) = (0, 0, \ldots, 1)'$. Transposing and rearranging this last expression, we get $\omega'\Omega^{-1}/\sigma_u^2 = \rho(0, 0, \ldots, 1)$. This means that the last term in (9.18) is equal to $\rho(0, 0, \ldots, 1)(y - X\widehat{\beta}_{GLS}) = \rho e_{T,GLS}$, where $e_{T,GLS}$ is the T-th GLS residual. This differs from the spherical model prediction in that next year's disturbance is not independent of the sample disturbances and hence, is not predicted by its mean which is zero. Instead, one uses the fact that $u_{T+1} = \rho u_T + \epsilon_{T+1}$ and predicts u_{T+1} by $\rho e_{T,GLS}$. Only ϵ_{T+1} is predicted by its zero mean but u_T is predicted by $e_{T,GLS}$.

9.7 Unknown Ω

If Ω is unknown, the practice is to get a consistent estimate of Ω, say $\widehat{\Omega}$ and substitute that in $\widehat{\beta}_{GLS}$. The resulting estimator is

$$\widehat{\beta}_{FGLS} = (X'\widehat{\Omega}^{-1}X)^{-1}X'\widehat{\Omega}^{-1}y \tag{9.19}$$

and is called a feasible GLS estimator of β. Once $\widehat{\Omega}$ replaces Ω the Gauss-Markov Theorem no longer necessarily holds. In other words, $\widehat{\beta}_{FGLS}$ is not BLUE, although it is still consistent.

The finite sample properties of $\widehat{\beta}_{FGLS}$ are in general difficult to derive. However, we have the following asymptotic results.

Theorem 1: $\sqrt{n}(\widehat{\beta}_{GLS} - \beta)$ and $\sqrt{n}(\widehat{\beta}_{FGLS} - \beta)$ have the *same asymptotic distribution* $N(0, \sigma^2 Q^{-1})$, where $Q = \lim(X'\Omega^{-1}X)/n$ as $n \to \infty$, if (i) plim $X'(\widehat{\Omega}^{-1} - \Omega^{-1})X/n = 0$ and (ii) plim $X'(\widehat{\Omega}^{-1} - \Omega^{-1})u/n = 0$. A sufficient condition for this theorem to hold is that $\widehat{\Omega}$ is a consistent estimator of Ω and X has a satisfactory limiting behavior.

Lemma 1: If in addition plim $u'(\widehat{\Omega}^{-1} - \Omega^{-1})u/n = 0$, then $s^{*2} = e'_{GLS}\Omega^{-1}e_{GLS}/(n-K)$ and $\widehat{s}^{*2} = e'_{FGLS}\widehat{\Omega}^{-1}e_{FGLS}/(n-K)$ are both *consistent* for σ^2. This means that one can perform test of hypotheses based on asymptotic arguments using $\widehat{\beta}_{FGLS}$ and $\widehat{s}^{*2}$ rather than $\widehat{\beta}_{GLS}$ and s^{*2}, respectively. For a proof of Theorem 1 and Lemma 1, see Theil (1971), Schmidt (1976) or Judge et al. (1985).

Monte Carlo evidence under heteroskedasticity or serial correlation suggest that there is gain in performing feasible GLS rather than OLS in finite samples. However, we have also seen in Chapter 5 that performing a two-step Cochrane-Orcutt procedure is not necessarily better than OLS if the X's are trended. This says that feasible GLS omitting the first observation (in this case Cochrane-Orcutt) may not be better in finite samples than OLS using all the observations.

9.8 The W, LR and LM Statistics Revisited

In this section we present a simplified and more general proof of $W \geq LR \geq LM$ due to Breusch (1979). For the general linear model given in (9.1) with $u \sim N(0, \Sigma)$ and H_0; $R\beta = r$. The likelihood function given in (9.14) with $\Sigma = \sigma^2\Omega$, can be maximized with respect to β and Σ without imposing H_0, yielding the unrestricted estimators $\widehat{\beta}_u$ and $\widehat{\Sigma}$, where $\widehat{\beta}_u = (X'\widehat{\Sigma}^{-1}X)^{-1}X'\widehat{\Sigma}^{-1}y$. Similarly, this likelihood can be maximized subject to the restriction H_0, yielding $\widetilde{\beta}_r$ and $\widetilde{\Sigma}$, where

$$\widetilde{\beta}_r = (X'\widetilde{\Sigma}^{-1}X)^{-1}X'\widetilde{\Sigma}^{-1}y - (X'\widetilde{\Sigma}^{-1}X)^{-1}R'\widetilde{\mu} \tag{9.20}$$

as in (9.17), where $\widetilde{\mu} = \widetilde{A}^{-1}(R\widehat{\beta}_r - r)$ is the Lagrange multiplier described in equation (7.35) of Chapter 7 and $\widetilde{A} = [R(X'\widetilde{\Sigma}^{-1}X)^{-1}R']$. The major distinction from Chapter 7 is that Σ is unknown and has to be estimated. Let $\widehat{\beta}_r$ denote the unrestricted maximum likelihood estimator of β conditional on the restricted variance-covariance estimator $\widetilde{\Sigma}$ and let $\widetilde{\beta}_u$ denote the restricted maximum likelihood of β (satisfying H_0) conditional on the unrestricted variance-covariance estimator $\widehat{\Sigma}$. More explicitly,

$$\widehat{\beta}_r = (X'\widetilde{\Sigma}^{-1}X)^{-1}X'\widetilde{\Sigma}^{-1}y \tag{9.21}$$

and

$$\widetilde{\beta}_u = \widehat{\beta}_u - (X'\widehat{\Sigma}^{-1}X)^{-1}R'\widehat{A}^{-1}(R\widehat{\beta}_u - r) \tag{9.22}$$

Knowing Σ, the Likelihood Ratio statistic is given by

$$\begin{aligned} LR &= -2\log[\max_{R\beta=r} L(\beta/\Sigma)/\max_{\beta} L(\beta/\Sigma)] = -2\log[L(\widetilde{\beta},\Sigma)/L(\widehat{\beta},\Sigma)] \\ &= \widetilde{u}'\Sigma^{-1}\widetilde{u} - \widehat{u}'\Sigma^{-1}\widehat{u} \end{aligned} \tag{9.23}$$

where $\widehat{u} = y - X\widehat{\beta}$ and $\widetilde{u} = y - X\widetilde{\beta}$, both estimators of β are conditional on a *known* Σ.

$$R\widehat{\beta}_u \sim N(R\beta, R(X'\widehat{\Sigma}^{-1}X)^{-1}R')$$

and the Wald statistic is given by

$$W = (R\widehat{\beta}_u - r)'\widehat{A}^{-1}(R\widehat{\beta}_u - r) \quad \text{where} \quad \widehat{A} = [R(X'\widehat{\Sigma}^{-1}X)^{-1}R'] \tag{9.24}$$

Using (9.22), it is easy to show that $\widetilde{u}_u = y - X\widetilde{\beta}_u$ and $\widehat{u}_u = y - X\widehat{\beta}_u$ are related as follows:

$$\widetilde{u}_u = \widehat{u}_u + X(X'\widehat{\Sigma}^{-1}X)^{-1}R'\widehat{A}^{-1}(R\widehat{\beta}_u - r) \tag{9.25}$$

and

$$\widetilde{u}_u'\widehat{\Sigma}^{-1}\widetilde{u}_u = \widehat{u}_u'\widehat{\Sigma}^{-1}\widehat{u}_u + (R\widehat{\beta}_u - r)'\widehat{A}^{-1}(R\widehat{\beta}_u - r) \tag{9.26}$$

The cross-product terms are zero because $X'\widehat{\Sigma}^{-1}\widehat{u}_u = 0$. Therefore,

$$\begin{aligned} W &= \widetilde{u}_u'\widehat{\Sigma}^{-1}\widetilde{u}_u - \widehat{u}_u'\widehat{\Sigma}^{-1}\widehat{u}_u = -2\log[L(\widetilde{\beta},\widehat{\Sigma})/L(\widehat{\beta},\widehat{\Sigma})] \\ &= -2\log[\max_{R\beta=r} L(\beta/\widehat{\Sigma})/\max_{\beta} L(\beta/\widehat{\Sigma})] \end{aligned} \tag{9.27}$$

and the Wald statistic can be interpreted as a LR statistic conditional on $\widehat{\Sigma}$, the unrestricted maximum likelihood estimator of Σ.

Similarly, the Lagrange multiplier statistic, which tests that $\mu = 0$, is given by

$$LM = \mu'\widetilde{A}\mu = (R\widehat{\beta}_r - r)'\widetilde{A}^{-1}(R\widehat{\beta}_r - r) \tag{9.28}$$

Using (9.20) one can easily show that

$$\widetilde{u}_r = \widehat{u}_r + X(X'\widetilde{\Sigma}^{-1}X)^{-1}R'\widetilde{A}^{-1}(R\widehat{\beta}_r - r) \tag{9.29}$$

and

$$\widetilde{u}_r'\widetilde{\Sigma}^{-1}\widetilde{u}_r = \widehat{u}_r'\widetilde{\Sigma}^{-1}\widehat{u}_r + \mu'\widetilde{A}\mu \tag{9.30}$$

The cross-product terms are zero because $X'\widetilde{\Sigma}^{-1}\widehat{u}_r = 0$. Therefore,

$$\begin{aligned} LM &= \widetilde{u}_r'\widetilde{\Sigma}^{-1}\widetilde{u}_r - \widehat{u}_r'\widetilde{\Sigma}^{-1}\widehat{u}_r = -2\log[L(\widetilde{\beta}_r,\widetilde{\Sigma})/L(\widehat{\beta}_r,\widetilde{\Sigma})] \\ &= -2\log[\max_{R\beta=r} L(\beta/\widetilde{\Sigma})/\max_{\beta} L(\beta/\widetilde{\Sigma})] \end{aligned} \tag{9.31}$$

and the Lagrange multiplier statistic can be interpreted as a LR statistic conditional on $\widetilde{\Sigma}$ the restricted maximum likelihood of Σ. Given that

$$\max_{\beta} L(\beta/\widetilde{\Sigma}) \leq \max_{\beta,\Sigma} L(\beta,\Sigma) = \max_{\beta} L(\beta/\widehat{\Sigma}) \tag{9.32}$$

$$\max_{R\beta=r} L(\beta/\widehat{\Sigma}) \leq \max_{R\beta=r,\Sigma} L(\beta,\Sigma) = \max_{R\beta=r} L(\beta/\widetilde{\Sigma}) \tag{9.33}$$

it can be easily shown that the likelihood ratio statistic given by

$$LR = -2\log[\max_{R\beta=r,\Sigma} L(\beta,\Sigma)/\max_{\beta,\Sigma} L(\beta,\Sigma)] \tag{9.34}$$

satisfies the following inequality

$$W \geq LR \geq LM \tag{9.35}$$

The proof is left to the reader, see problem 6.

This general and simple proof holds as long as the maximum likelihood estimator of β is uncorrelated with the maximum likelihood estimator of Σ, see Breusch (1979).

9.9 Spatial Error Correlation[1]

Unlike time-series, there is typically no unique natural ordering for cross-sectional data. Spatial autocorrelation permit correlation of the disturbance terms across cross-sectional units. There is an extensive literature on spatial models in regional science, urban economics, geography and statistics, see Anselin (1988). Examples in economics usually involve spillover effects or externalities due to geographical proximity. For example, the productivity of public capital, like roads and highways, on the output of neighboring states. Also, the pricing of welfare in one state that pushes recipients to other states. Spatial correlation could relate directly to the model dependent variable y, the exogenous variables X, the disturbance term u, or to a combination of all three. Here we consider spatial correlation in the disturbances and leave the remaining literature on spatial dependence to the motivated reader to pursue in Anselin (1988, 2001) and Anselin and Bera (1998) to mention a few.

For the cross-sectional disturbances, the spatial autocorrelation is specified as

$$u = \lambda W u + \epsilon \tag{9.36}$$

where λ is the spatial autoregressive coefficient satisfying $|\lambda| < 1$, and $\epsilon \sim \text{IIN}(0, \sigma^2)$. W is a *known* spatial weight matrix with diagonal elements equal to zero. W also satisfies some other regularity conditions like the fact that $I_n - \lambda W$ must be nonsingular.

The regression model given in (9.1) can be written as

$$y = X\beta + (I_n - \lambda W)^{-1}\epsilon \tag{9.37}$$

with the variance-covariance matrix of the disturbances given by

$$\Sigma = \sigma^2\Omega = \sigma^2(I_n - \lambda W)^{-1}(I_n - \lambda W')^{-1} \tag{9.38}$$

Under normality of the disturbances, Ord (1975) derived the maximum likelihood estimators

$$\ln L = -\frac{1}{2}\ln|\Omega| - \frac{n}{2}\ln 2\pi\sigma^2 - (y - X\beta)'\Omega^{-1}(y - X\beta)/2\sigma^2 \tag{9.39}$$

The Jacobian term simplifies by using

$$\ln|\Omega| = -2\ln|I - \lambda W| = -2\sum_{i=1}^{n}\ln(1 - \lambda w_i) \tag{9.40}$$

where w_i are the eigenvalues of the spatial weight matrix W. The first-order conditions yield the familiar GLS estimator of β and the associated estimator of σ^2:

$$\widehat{\beta}_{MLE} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \quad \text{and} \quad \widehat{\sigma}^2_{MLE} = e'_{MLE}\Omega^{-1}e_{MLE}/n \tag{9.41}$$

where $e_{MLE} = y - X\widehat{\beta}_{MLE}$. An estimate of λ can be obtained using the iterative solution of the first-order conditions in Magnus (1978, p. 283):

$$-\frac{1}{2}\,\text{tr}\left[\left(\frac{\partial\Omega^{-1}}{\partial\lambda}\right)\Omega\right] = e'_{MLE}\left(\frac{\partial\Omega^{-1}}{\partial\lambda}\right)e_{MLE} \tag{9.42}$$

where

$$\partial\Omega^{-1}/\partial\lambda = -W - W' + \lambda W'W \tag{9.43}$$

Alternatively, one can substitute $\widehat{\beta}_{MLE}$ and $\widehat{\sigma}^2_{MLE}$ from (9.41) into the log-likelihood in (9.39) to get the concentrated log-likelihood which will be a nonlinear function of λ, see Anselin (1988) for details.

Testing for zero spatial autocorrelation i.e., H_0; $\lambda = 0$ is usually based on the Moran I-test which is similar to the Durbin-Watson statistic in time-series. This is given by

$$MI = \frac{n}{S_0}\left(\frac{e'We}{e'e}\right) \tag{9.44}$$

where e denotes the vector of OLS residuals and S_0 is a standardization factor equal to the sum of the spatial weights $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$. For a row-standardized weights matrix W where each row sums to one, $S_0 = n$ and the Moran I-statistic simplifies to $e'We/e'e$. In practice the test is implemented by standardizing it and using the asymptotic $N(0,1)$ critical values, see Anselin and Bera (1988). In fact, for a row-standardized W matrix, the mean and variance of the Moran I-statistic is obtained from

$$E(MI) = E\left(\frac{e'We}{e'e}\right) = \mathrm{tr}(\bar{P}_X W)/(n-k) \tag{9.45}$$

and

$$E(MI)^2 = \frac{\mathrm{tr}(\bar{P}_X W \bar{P}_X W') + \mathrm{tr}(\bar{P}_X W)^2 + \{\mathrm{tr}(\bar{P}_{\bar{X}} W)\}^2}{(n-k)(n-k+2)}$$

Alternatively, one can derive the Lagrange Multiplier test for H_0; $\lambda = 0$ using the result that $\partial \ln L/\partial\lambda$ evaluated under the null of $\lambda = 0$ is equal to $u'Wu/\sigma^2$ and the fact that the Information matrix is block-diagonal between β and (σ^2, λ), see problem 14. In fact, one can show that

$$LM_\lambda = \frac{(e'We/\widetilde{\sigma}^2)^2}{\mathrm{tr}[(W'+W)W]} \tag{9.46}$$

with $\widetilde{\sigma}^2 = e'e/n$. Under H_0, LM_λ is asymptotically distributed as χ^2_1. One can clearly see the connection between Moran's I-statistic and LM_λ. Computationally, the W and LR tests are more demanding since the require ML estimation under spatial autocorrelation.

This is only a brief introduction into the spatial dependence literature. Hopefully, it will motivate the reader to explore alternative formulations of spatial dependence, alternative estimation and testing methods discussed in this literature and the numerous applications in economics on hedonic housing, crime rates, police expenditures and R&D spillovers, to mention a few.

Note

1. This section is based on Anselin (1988, 2001) and Anselin and Bera (1998).

Problems

1. (a) Using equation (7.5) of Chapter 7, verify that $\text{var}(\widehat{\beta}_{OLS})$ is that given in (9.5).

   (b) Show that $\text{var}(\widehat{\beta}_{OLS}) - \text{var}(\widehat{\beta}_{GLS}) = \sigma^2 A\Omega A'$ where

   $$A = [(X'X)^{-1}X' - (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}].$$

   Conclude that this difference in variances is positive semi-definite.

2. (a) Show that $E(s^2) = \sigma^2 \text{tr}(\Omega \bar{P}_X)/(n-K) \neq \sigma^2$. **Hint:** Follow the same proof given below equation (7.6) of Chapter 7, but substitute $\sigma^2\Omega$ instead of $\sigma^2 I_n$.

   (b) Use the fact that P_X and Σ are non-negative definite matrices with $\text{tr}(\Sigma P_X) \geq 0$ to show that $0 \leq E(s^2) \leq \text{tr}(\Sigma)/(n-K)$ where $\text{tr}(\Sigma) = \sum_{i=1}^{n} \sigma_i^2$ with $\sigma_i^2 = \text{var}(u_i) \geq 0$. This bound was derived by Dufour (1986). Under homoskedasticity, show that this bound becomes $0 \leq E(s^2) \leq n\sigma^2/(n-K)$. In general, $0 \leq \{\text{mean of } n-K \text{ smallest characteristic roots of } \Sigma\} \leq E(s^2) \leq \{\text{mean of } n-K \text{ largest characteristic roots of } \Sigma\} \leq \text{tr}(\Sigma)/(n-K)$, see Sathe and Vinod (1974) and Neudecker (1977, 1978).

   (c) Show that a *sufficient condition* for s^2 to be consistent for σ^2 irrespective of X is that λ_{max} = the largest characteristic root of Ω is $o(n)$, i.e., $\lambda_{max}/n \to 0$ as $n \to \infty$ *and* plim $(u'u/n) = \sigma^2$. **Hint:** $s^2 = u'\bar{P}_X u/(n-K) = u'u/(n-K) - u'P_X u/(n-K)$. By assumption, the first term tends in probability limits to σ^2 as $n \to \infty$. The second term has expectation $\sigma^2 \text{tr}(P_X\Omega)/(n-K)$. Now $P_X\Omega$ has rank K and therefore exactly K non-zero characteristic roots each of which cannot exceed λ_{max}. This means that $E[u'P_X u/(n-K)] \leq \sigma^2 K\lambda_{max}/(n-K)$. Using the condition that $\lambda_{max}/n \to 0$ proves the result. See Krämer and Berghoff (1991).

   (d) Using the same reasoning in part (a), show that s^{*2} given in (9.6) is unbiased for σ^2.

3. *The AR(1) Model.* See Kadiyala (1968).

   (a) Verify that $\Omega\Omega^{-1} = I_T$ for Ω and Ω^{-1} given in (9.9) and (9.10), respectively.

   (b) Show that $P^{-1\prime}P^{-1} = (1-\rho^2)\Omega^{-1}$ for P^{-1} defined in (9.11).

   (c) Conclude that $\text{var}(P^{-1}u) = \sigma_\epsilon^2 I_T$. **Hint:** $\Omega = (1-\rho^2)PP'$ as can be easily derived from part (b).

4. *Restricted GLS.* Using the derivation of the restricted least squares estimator for $u \sim (0, \sigma^2 I_n)$ in Chapter 7, verify equation (9.17) for the restricted GLS estimator based on $u \sim (0, \sigma^2\Omega)$. **Hint:** Apply restricted least squares results to the transformed model given in (9.3).

5. *Best Linear Unbiased Prediction.* This is based on Goldberger (1962). Consider all linear predictors of $y_{T+s} = x'_{T+s}\beta + u_{T+s}$ of the form $\widehat{y}_{T+s} = c'y$, where $u \sim (0, \Sigma)$ and $\Sigma = \sigma^2\Omega$.

   (a) Show that $c'X = x'_{T+s}$ for $\widehat{y}_{T+s}$ to be unbiased.

   (b) Show that $\text{var}(\widehat{y}_{T+s}) = c'\Sigma c + \sigma^2_{T+s} - 2c'\omega$ where $\text{var}(u_{T+s}) = \sigma^2_{T+s}$ and $\omega = E(u_{T+s}u)$.

   (c) Minimize $\text{var}(\widehat{y}_{T+s})$ given in part (b) subject to $c'X = x'_{T+s}$ and show that

   $$\widehat{c} = \Sigma^{-1}[I_T - X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}]\omega + \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}x_{T+s}$$

   This means that $\widehat{y}_{T+s} = \widehat{c}'y = x'_{T+s}\widehat{\beta}_{GLS} + \omega'\Sigma^{-1}e_{GLS} = x'_{T+s}\widehat{\beta}_{GLS} + \omega'\Omega^{-1}e_{GLS}/\sigma^2$. For $s=1$, i.e., predicting one period ahead, this verifies equation (9.18). **Hint:** Use partitioned inverse in solving the first-order minimization equations.

   (d) Show that $\widehat{y}_{T+s} = x'_{T+s}\widehat{\beta}_{GLS} + \rho^s e_{T,GLS}$ for the stationary AR(1) disturbances with autoregressive parameter ρ, and $|\rho| < 1$.

6. *The W, LR and LM Inequality.* Using the inequalities given in equations (9.32) and (9.33) verify equation (9.35) which states that $W \geq LR \geq LM$. **Hint:** Use the conditional likelihood ratio interpretations of W and LM given in equations (9.27) and (9.31) respectively.

7. Consider the simple linear regression

$$y_i = \alpha + \beta X_i + u_i \quad i = 1, 2, \ldots, n$$

with $u_i \sim \text{IIN}(0, \sigma^2)$. For H_0; $\beta = 0$, derive the LR, W and LM statistics in terms of conditional likelihood ratios as described in Breusch (1979). In other words, compute $W = -2\log[\max_{H_0}(\alpha, \beta/\widetilde{\sigma}^2)/\max_{\alpha,\beta} L(\alpha, \beta/\widehat{\sigma}^2)]$, $LM = -2\log[\max_{H_0} L(\alpha, \beta/\widetilde{\sigma}^2)/\max_{\alpha,\beta} L(\alpha, \beta/\widetilde{\sigma}^2)]$ and $LR = -2\log[\max_{H_0} L(\alpha, \beta, \sigma^2)/\max_{\alpha,\beta,\sigma^2} L(\alpha, \beta, \sigma^2)]$ where $\widehat{\sigma}^2$ is the unrestricted MLE of σ^2 while $\widetilde{\sigma}^2$ is the restricted MLE of σ^2 under H_0. Use these results to infer that $W \geq LR \geq LM$.

8. *Sampling Distributions and Efficiency Comparison of OLS and GLS.* Consider the following regression model $y_t = \beta x_t + u_t$ for $(t = 1, 2)$, where $\beta = 2$ and x_t takes on the fixed values $x_1 = 1$, $x_2 = 2$. The u_t's have the following discrete joint probability distribution:

(u_1, u_2)	Probability
$(-1, -2)$	1/8
$(1, -2)$	3/8
$(-1, 2)$	3/8
$(1, 2)$	1/8

   (a) What is the variance-covariance matrix of the disturbances? Are the disturbances heteroskedastic? Are they correlated?

   (b) Find the sampling distributions of $\widehat{\beta}_{OLS}$ and $\widetilde{\beta}_{GLS}$ and verify that $\text{var}(\widehat{\beta}_{OLS}) > \text{var}(\widetilde{\beta}_{GLS})$.

   (c) Find the sampling distribution of the OLS residuals and verify that the estimated $\text{var}(\widehat{\beta}_{OLS})$ is biased. Also, find the sampling distribution of the GLS residuals and verify that the MSE of the GLS regression is an unbiased estimator of the GLS regression variance. **Hint:** Read Oksanen (1991) and Phillips and Wickens (1978), pp. 3-4. This problem is based on Baltagi (1992). See also the solution by Im and Snow (1993).

9. *Equi-correlation.* This problem is based on Baltagi (1998). Consider the regression model given in (9.1) with *equi-correlated* disturbances, i.e., equal variances and equal covariances: $E(u'u) = \sigma^2\Omega = \sigma^2[(1-\rho)I_T + \rho\iota_T\iota_T']$ where ι_T is a vector of ones of dimension T and I_T is the identity matrix. In this case, $\text{var}(u_t) = \sigma^2$ and $\text{cov}(u_t, u_s) = \rho\sigma^2$ for $t \neq s$ with $t = 1, 2, \ldots, T$. Assume that the regression has a constant.

   (a) Show that OLS on this model is equivalent to GLS. **Hint:** Verify Zyskind's condition given in (9.8) using the fact that $P_X\iota_T = \iota_T$ if ι_T is a column of X.

   (b) Show that $E(s^2) = \sigma^2(1-\rho)$. Also, that Ω is positive semi-definite when $-1/(T-1) \leq \rho \leq 1$. Conclude that if $-1/(T-1) \leq \rho \leq 1$, then $0 \leq E(s^2) \leq [T/(T-1)]\sigma^2$. The lower and upper bounds are attained at $\rho = 1$ and $\rho = -1/(T-1)$, respectively, see Dufour (1986). **Hint:** Ω is positive semi-definite if for every arbitrary non-zero vector a we have $a'\Omega a \geq 0$. What is this expression for $a = \iota_T$?

   (c) Show that for this equi-correlated regression model, the BLUP of $y_{T+1} = x'_{T+1}\beta + u_{T+1}$ is $\widehat{y}_{T+1} = x'_{T+1}\widehat{\beta}_{OLS}$ as long as there is a constant in the model.

10. Consider the simple regression with no regressors and equi-correlated disturbances:

$$y_i = \alpha + u_i \quad i = 1, \ldots, n$$

where $E(u_i) = 0$ and

$$\begin{aligned} \text{cov}(u_i, u_j) &= \rho\sigma^2 \quad \text{for} \quad i \neq j \\ &= \sigma^2 \quad \text{for} \quad i = j \end{aligned}$$

with $\frac{1}{(n-1)} \leq \rho \leq 1$ for the variance-covariance matrix of the disturbances to be positive definite.

(a) Show that the OLS and GLS estimates of α are identical. This is based on Kruskal (1968).

(b) Show that the bias in s^2, the OLS estimator of σ^2, is given by $-\rho\sigma^2$.

(c) Show that the GLS estimator of σ^2 is unbiased.

(d) Show that the E[estimated var$(\widehat{\alpha})-$ true var$(\widehat{\alpha}_{OLS})$] is also $-\rho\sigma^2$.

11. *Prediction Error Variances Under Heteroskedasticity.* This is based on Termayne (1985). Consider the t-th observation of the linear regression model given in (9.1).

$$y_t = x_t'\beta + u_t \qquad t = 1, 2, \ldots, T$$

where y_t is a scalar x_t' is $1 \times K$ and β is a $K \times 1$ vector of unknown coefficients. u_t is assumed to have zero mean, heteroskedastic variances $E(u_t^2) = (z_t'\gamma)^2$ where z_t' is a $1 \times r$ vector of observed variables and γ is an $r \times 1$ vector of parameters. Furthermore, these u_t's are not serially correlated, so that $E(u_t u_s) = 0$ for $t \neq s$.

(a) Find the var$(\widehat{\beta}_{OLS})$ and var$(\widetilde{\beta}_{GLS})$ for this model.

(b) Suppose we are forecasting y for period f in the future knowing x_f, i.e., $y_f = x_f'\beta + u_f$ with $f > T$. Let $\widehat{e}_f$ and $\widetilde{e}_f$ be the forecast errors derived using OLS and GLS, respectively. Show that the prediction error variances of the point predictions of y_f are given by

$$\text{var}(\widehat{e}_f) = x_f'(\textstyle\sum_{t=1}^{T} x_t x_t')^{-1}[\sum_{t=1}^{T} x_t x_t'(z_t'\gamma)^2](\sum_{t=1}^{T} x_t x_t')^{-1}x_f + (z_f'\gamma)^2$$

$$\text{var}(\widetilde{e}_f) = x_f'[\textstyle\sum_{t=1}^{T} x_t x_t'(z_t'\gamma)^2]^{-1}x_f + (z_f'\gamma)^2$$

(c) Show that the variances of the two forecast errors of conditional mean $E(y_f/x_f)$ based upon $\widehat{\beta}_{OLS}$ and $\widetilde{\beta}_{GLS}$ and denoted by $\widehat{c}_f$ and $\widetilde{c}_f$, respectively are the first two terms of the corresponding expressions in part (b).

(d) Now assume that $K = 1$ and $r = 1$ so that there is only one single regressor x_t and one z_t variable determining the heteroskedasticity. Assume also for simplicity that the empirical moments of x_t match the population moments of a Normal random variable with mean zero and variance θ. Show that the relative efficiency of the OLS to the GLS predictor of y_f is equal to $(T+1)/(T+3)$, whereas the relative efficiency of the corresponding ratio involving the two predictions of the conditional mean is $(1/3)$.

12. *Estimation of Time Series Regressions with Autoregressive Disturbances and Missing Observations.* This is based on Baltagi and Wu (1997). Consider the following time series regression model,

$$y_t = x_t'\beta + u_t \quad t = 1, \ldots, T,$$

where β is a $K \times 1$ vector of regression coefficients including the intercept. The disturbances follow a stationary AR(1) process, that is,

$$u_t = \rho u_{t-1} + \epsilon_t,$$

with $|\rho| < 1$, ϵ_t is IIN$(0, \sigma_\epsilon^2)$, and $u_0 \sim N(0, \sigma_\epsilon^2/(1-\rho^2))$. This model is only observed at times t_j for $j = 1, \ldots, n$ with $1 = t_1 < \ldots < t_n = T$ and $n > K$. The typical covariance element of u_t for the observed periods t_j and t_s is given by

$$\text{cov}(u_{t_j}, u_{t_s}) = \frac{\sigma_\epsilon^2}{1-\rho^2}\ \rho^{|t_j - t_s|} \quad \text{for} \quad s, j = 1, \ldots, n$$

Knowing ρ, derive a simple Prais-Winsten-type transformation that will obtain GLS as a simple least squares regression.

13. *Multiplicative Heteroskedasticity.* This is based on Harvey (1976). Consider the linear model given in (9.1) and let $u \sim N(0, \Sigma)$ where $\Sigma = \text{diag}[\sigma_i^2]$. Assume that $\sigma_i^2 = \sigma^2 h_i(\theta)$ with $\theta' = (\theta_1, \ldots, \theta_s)$ and $h_i(\theta) = \exp(\theta_1 z_{1i} + \ldots + \theta_s z_{si}) = \exp(z_i'\theta)$ with $z_i' = (z_{1i}, \ldots, z_{si})$.

(a) Show that log-likelihood function is given by

$$\log L(\beta, \theta, \sigma^2) = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2}\sum_{i=1}^N \log h_i(\theta) - \frac{1}{2\sigma^2}\sum_{i=1}^N \frac{(y_i - x_i'\beta)^2}{h_i(\theta)}$$

and the score with respect to θ is

$$\partial \log L/\partial\theta = -\frac{1}{2}\sum_{i=1}^N \frac{1}{h_i(\theta)}\frac{\partial h_i}{\partial\theta} + \frac{1}{2\sigma^2}\sum_{i=1}^N \frac{(y_i - x_i'\beta)^2}{(h_i(\theta))^2}\cdot\frac{\partial h_i}{\partial\theta}$$

Conclude that for multiplicative heteroskedasticity, equating this score to zero yields

$$\sum_{i=1}^N \frac{(y_i - x_i'\beta)^2}{\exp(z_i'\theta)} z_i = \sigma^2 \sum_{i=1}^N z_i.$$

(b) Show that the Information matrix is given by

$$I(\beta, \theta, \sigma^2) = \begin{bmatrix} X'\Sigma^{-1}X & 0 & 0 \\ 0 & \frac{1}{2}\sum_{i=1}^N \frac{1}{(h_i(\theta))^2}\frac{\partial h_i}{\partial\theta}\frac{\partial h_i}{\partial\theta'} & \frac{1}{2\sigma^2}\sum_{i=1}^N \frac{1}{h_i(\theta)}\frac{\partial h_i}{\partial\theta} \\ 0 & \frac{1}{2\sigma^2}\sum_{i=1}^N \frac{1}{h_i(\theta)}\frac{\partial h_i}{\partial\theta'} & \frac{N}{2\sigma^4} \end{bmatrix}$$

and for multiplicative heteroskedasticity this becomes

$$I(\beta, \theta, \sigma^2) = \begin{bmatrix} X'\Sigma^{-1}X & 0 & 0 \\ 0 & \frac{1}{2}Z'Z & \frac{1}{2\sigma^2}\sum_{i=1}^N z_i \\ 0 & \frac{1}{2\sigma^2}\sum_{i=1}^N z_i' & \frac{N}{2\sigma^4} \end{bmatrix}$$

where $Z_i' = (z_1, \ldots, z_N)$.

(c) Assume that $h_i(\theta)$ satisfies $h_i(0) = 1$, then the test for heteroskedasticity is H_0; $\theta = 0$ versus H_1; $\theta \neq 0$. Show that the score with respect to θ and σ^2 evaluated under the null hypothesis, i.e., at $\theta = 0$ and $\widetilde{\sigma}^2 = e'e/N$ is given by

$$\widetilde{S} = \begin{pmatrix} \frac{1}{2}\sum_{i=1}^N z_i \left(\frac{e_i^2}{\widetilde{\sigma}^2} - 1\right) \\ 0 \end{pmatrix}$$

where e denotes the vector of OLS residuals. The Information matrix with respect to θ and σ^2can be obtained from the bottom right block of $I(\beta, \theta, \sigma^2)$ given in part (b). Conclude that the score test for H_0 is given by

$$LM = \frac{\sum_{i=1}^{N} z_i'(e_i^2 - \widetilde{\sigma}^2) \left(\sum_{i=1}^{N} (z_i - \bar{z}) (z_i - \bar{z})' \right)^{-1} \sum_{i=1}^{N} z_i (e_i^2 - \widetilde{\sigma}^2)}{2\widetilde{\sigma}^4}$$

This statistic is asymptotically distributed as χ_s^2 under H_0. From Chapter 5, we can see that this is a special case of the Breusch and Pagan (1979) test-statistic which can be obtained as one-half the regression sum of squares of $e^2/\widetilde{\sigma}^2$ on a constant and Z. Koenker and Bassett (1982) suggested replacing the denominator $2\widetilde{\sigma}^4$ by $\sum_{i=1}^{N} (e^2 - \widetilde{\sigma}^2)^2/N$ to make this test more robust to departures from normality.

14. *Spatial Autocorrelation.* Consider the regression model given in (9.1) with spatial autocorrelation defined in (9.36).

    (a) Verify that the first-order conditions of maximization of the log-likelihood function given in (9.39) yield (9.41).

    (b) Show that for testing H_0; $\lambda = 0$, the score $\partial \ln L/\partial \lambda$ evaluated under the null, i.e., at $\lambda = 0$, is given by $u'Wu/\sigma^2$.

    (c) Show that the Information matrix with respect to $\sigma^2$and λ, evaluated under the null of $\lambda = 0$, is given by

$$\begin{bmatrix} \frac{n}{2\sigma^4} & \frac{\mathrm{tr}(W)}{\sigma^2} \\ \frac{\mathrm{tr}(W)}{\sigma^2} & \mathrm{tr}(W^2) + \mathrm{tr}(W'W) \end{bmatrix}$$

    (d) Conclude from parts (b) and (c) that the Lagrange Multiplier for H_0; $\lambda = 0$ is given by LM_λ in (9.46). **Hint:** Use the fact that the diagonal elements of W are zero, hence $\mathrm{tr}(W) = 0$.

References

Additional readings on GLS can be found in the econometric texts cited in the Preface.

Anselin, L. (2001), "Spatial Econometrics," Chapter 14 in B.H. Baltagi (ed.) *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Anselin, L. (1988), *Spatial Econometrics: Methods and Models* (Kluwer: Dordrecht).

Anselin, L. and A.K. Bera (1998), "Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics," in A. Ullah and D.E.A. Giles (eds.) *Handbook of Applied Economic Statistics* (Marcel Dekker: New York).

Balestra, P. (1970), "On the Efficiency of Ordinary Least Squares in Regression Models," *Journal of the American Statistical Association*, 65: 1330-1337.

Balestra, P. (1980), "A Note on the Exact Transformation Associated with First-Order Moving Average Process," *Journal of Econometrics*, 14: 381-394.

Baltagi, B.H. (1989), "Applications of a Necessary and Sufficient Condition for OLS to be BLUE," *Statistics and Probability Letters*, 8: 457-461.

Baltagi, B.H. (1992), "Sampling Distributions and Efficiency Comparisons of OLS and GLS in the Presence of Both Serial Correlation and Heteroskedasticity," *Econometric Theory*, Problem 92.2.3, 8: 304-305.

Baltagi, B.H. and P.X. Wu (1997), "Estimation of Time Series Regressions with Autoregressive Disturbances and Missing Observations," *Econometric Theory*, Problem 97.5.1, 13: 889.

Baltagi, B.H. (1998), "Prediction in the Equicorrelated Regression Model," *Econometric Theory*, Problem 98.3.3, 14: 382.

Breusch, T.S. (1979), "Conflict Among Criteria for Testing Hypotheses: Extensions and Comments," *Econometrica*, 47: 203-207.

Breusch, T.S. and A.R. Pagan (1979), "A Simple Test for Heteroskedasticity and Random Coefficient Variation," *Econometrica*, 47: 1287-1294.

Buse, A. (1982), "The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note," *The American Statistician*, 36: 153-157.

Dufour, J.M. (1986), "Bias of s^2 in Linear Regressions with Dependent Errors," *The American Statistician*, 40: 284-285.

Fuller, W.A. and G.E. Battese (1974), "Estimation of Linear Models with Crossed-Error Structure," *Journal of Econometrics*, 2: 67-78.

Goldberger, A.S. (1962), "Best Linear Unbiased Prediction in the Generalized Linear Regression Model," *Journal of the American Statistical Association*, 57: 369-375.

Harvey, A.C. (1976), "Estimating Regression Models With Multiplicative Heteroskedasticity," *Econometrica*, 44: 461-466.

Im, E.I. and M.S. Snow (1993), "Sampling Distributions and Efficiency Comparisons of OLS and GLS in the Presence of Both Serial Correlation and Heteroskedasticity," *Econometric Theory*, Solution 92.2.3, 9: 322-323.

Kadiyala, K.R. (1968), "A Transformation Used to Circumvent the Problem of Autocorrelation," *Econometrica*, 36: 93-96.

Koenker, R. and G. Bassett, Jr. (1982), "Robust Tests for Heteroskedasticity Based on Regression Quantiles," *Econometrica*, 50: 43-61.

Krämer, W. and S. Berghoff (1991), "Consistency of s^2 in the Linear Regression Model with Correlated Errors," *Empirical Economics*, 16: 375-377.

Kruskal, W. (1968), "When are Gauss-Markov and Least Squares Estimators Identical? A Coordinate-Free Approach," *The Annals of Mathematical Statistics*, 39: 70-75.

Lempers, F.B. and T. Kloek (1973), "On a Simple Transformation for Second-Order Autocorrelated Disturbances in Regression Analysis," *Statistica Neerlandica*, 27: 69-75.

Magnus, J. (1978), "Maximum Likelihood Estimation of the GLS Model with Unknown Parameters in the Disturbance Covariance Matrix," *Journal of Econometrics*, 7: 281-312.

Milliken, G.A. and M. Albohali (1984), "On Necessary and Sufficient Conditions for Ordinary Least Squares Estimators to be Best Linear Unbiased Estimators," *The American Statistician*, 38: 298-299.

Neudecker, H. (1977), "Bounds for the Bias of the Least Squares Estimator of σ^2 in Case of a First-Order Autoregressive Process (positive autocorrelation)," *Econometrica*, 45: 1257-1262.

Neudecker, H. (1978), "Bounds for the Bias of the LS Estimator in the Case of a First-Order (positive) Autoregressive Process Where the Regression Contains a Constant Term," *Econometrica*, 46: 1223-1226.

Newey, W. and K. West (1987), "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55: 703-708.

Oksanen, E.H. (1991), "A Simple Approach to Teaching Generalized Least Squares Theory," *The American Statistician*, 45: 229-233.

Ord, J.K. (1975), " Estimation Methods for Models of Spatial Interaction," *Journal of the American Statistical Association*, 70: 120-126.

Phillips, P.C.B. and M.R. Wickens (1978), *Exercises in Econometrics*, Vol. 1 (Philip Allan/Ballinger: Oxford).

Puntanen S. and G.P.H. Styan (1989), "The Equality of the Ordinary Least Squares Estimator and the Best Linear Unbiased Estimator," (with discussion), *The American Statistician*, 43: 153-161.

Sathe, S.T. and H.D. Vinod (1974), "Bounds on the Variance of Regression Coefficients Due to Heteroskedastic or Autoregressive Errors," *Econometrica*, 42: 333-340.

Schmidt, P. (1976), *Econometrics* (Marcell-Decker: New York).

Termayne, A.R. (1985), "Prediction Error Variances Under Heteroskedasticity," *Econometric Theory*, Problem 85.2.3, 1: 293-294.

Theil, H. (1971), *Principles of Econometrics* (Wiley: New York).

Thomas, J.J and K.F. Wallis (1971), "Seasonal Variation in Regression Analysis," *Journal of the Royal Statistical Society*, Series A, 134: 67-72.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48: 817-838.

Zyskind, G. (1967), "On Canonical Forms, Non-Negative Covariance Matrices and Best and Simple Least Squares Linear Estimators in Linear Models," *The Annals of Mathematical Statistics*, 38: 1092-1109.

CHAPTER 10
Seemingly Unrelated Regressions

When asked "How did you get the idea for SUR?" Zellner responded: "On a rainy night in Seattle in about 1956 or 1957, I somehow got the idea of algebraically writing a multivariate regression model in single equation form. When I figured out how to do that, everything fell into place because then many univariate results could be carried over to apply to the multivariate system and the analysis of the multivariate system is much simplified notationally, algebraically and, conceptually." Read the interview of Professor Arnold Zellner by Rossi (1989, p. 292).

10.1 Introduction

Consider two regression equations corresponding to two different firms

$$y_i = X_i\beta_i + u_i \qquad i = 1, 2 \tag{10.1}$$

where y_i and u_i are $T \times 1$ and X_i is $(T \times K_i)$ with $u_i \sim (0, \sigma_{ii} I_T)$. OLS is BLUE on each equation separately. Zellner's (1962) idea is to combine these Seemingly Unrelated Regressions in one stacked model, i.e.,

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \tag{10.2}$$

which can be written as

$$y = X\beta + u \tag{10.3}$$

where $y' = (y_1', y_2')$ and X and u are obtained similarly from (10.2). y and u are $2T \times 1$, X is $2T \times (K_1 + K_2)$ and β is $(K_1 + K_2) \times 1$. The stacked disturbances have a variance-covariance matrix

$$\Omega = \begin{bmatrix} \sigma_{11} I_T & \sigma_{12} I_T \\ \sigma_{21} I_T & \sigma_{22} I_T \end{bmatrix} = \Sigma \otimes I_T \tag{10.4}$$

where $\Sigma = [\sigma_{ij}]$ for $i, j = 1, 2$; with $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$ measuring the extent of correlation between the two regression equations. The Kronecker product operator $\otimes$ is defined in the Appendix to Chapter 7. Some important applications of SUR models in economics include the estimation of a system of demand equations or a translog cost function along with its share equations, see Berndt (1991). Briefly, a system of demand equations explains household consumption of several commodities. The correlation among equations could be due to unobservable household specific attributes that influence the consumption of these commodities. Similarly, in estimating a cost equation along with the corresponding input share equations based on firm level data. The correlation among equations could be due to unobservable firm-specific effects that influence input choice and cost in production decisions.

Problem 1 asks the reader to verify that OLS on the system of two equations in (10.2) yields the same estimates as OLS on each equation in (10.1) taken separately. If ρ is large we expect gain in efficiency in performing GLS rather than OLS on (10.3). In this case

$$\widehat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \tag{10.5}$$

where $\Omega^{-1} = \Sigma^{-1} \otimes I_T$. GLS will be BLUE for the system of two equations estimated jointly. Note that we only need to invert Σ to obtain Ω^{-1}. Σ is of dimension 2×2 whereas, Ω is of dimension $2T \times 2T$. In fact, if we denote by $\Sigma^{-1} = [\sigma^{ij}]$, then

$$\widehat{\beta}_{GLS} = \begin{bmatrix} \sigma^{11}X_1'X_1 & \sigma^{12}X_1'X_2 \\ \sigma^{21}X_2'X_1 & \sigma^{22}X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} \sigma^{11}X_1'y_1 + \sigma^{12}X_1'y_2 \\ \sigma^{21}X_2'y_1 + \sigma^{22}X_2'y_2 \end{bmatrix} \tag{10.6}$$

Zellner (1962) gave two sufficient conditions where it does not pay to perform GLS, i.e., GLS on this system of equations turns out to be OLS on each equation separately. These are the following:

Case 1: *Zero correlation* among the disturbances of the i-th and j-th equations, i.e., $\sigma_{ij} = 0$ for $i \neq j$. This means that Σ is diagonal which in turn implies that Σ^{-1} is diagonal with $\sigma^{ii} = 1/\sigma_{ii}$ for $i = 1, 2$, and $\sigma^{ij} = 0$ for $i \neq j$. Therefore, (10.6) reduces to

$$\widehat{\beta}_{GLS} = \begin{bmatrix} \sigma_{11}(X_1'X_1)^{-1} & 0 \\ 0 & \sigma_{22}(X_2'X_2)^{-1} \end{bmatrix} \begin{bmatrix} X_1'y_1/\sigma_{11} \\ X_2'y_2/\sigma_{22} \end{bmatrix} = \begin{bmatrix} \widehat{\beta}_{1,OLS} \\ \widehat{\beta}_{2,OLS} \end{bmatrix} \tag{10.7}$$

Case 2: *Same regressors* across all equations. This means that all the X_i's are the same, i.e., $X_1 = X_2 = X^*$. This rules out different number of regressors in each equation and all the X_i's must have the same dimension, i.e., $K_1 = K_2 = K$. Hence, $X = I_2 \otimes X^*$ and (10.6) reduces to

$$\begin{aligned} \widehat{\beta}_{GLS} &= [(I_2 \otimes X^{*\prime})(\Sigma^{-1} \otimes I_T)(I_2 \otimes X^*)]^{-1}[(I_2 \otimes X^{*\prime})(\Sigma^{-1} \otimes I_T)y] \\ &= [\Sigma \otimes (X^{*\prime}X^*)^{-1}][(\Sigma^{-1} \otimes X^{*\prime})y] = [I_2 \otimes (X^{*\prime}X^*)^{-1}X^{*\prime}]y = \widehat{\beta}_{OLS} \end{aligned} \tag{10.8}$$

These results generalize to the case of M regression equations, but for simplicity of exposition we considered the case of two equations only.

A necessary and sufficient condition for SUR(GLS) to be equivalent to OLS, was derived by Dwivedi and Srivastava (1978). An alternative derivation based on the Milliken and Albohali (1984) necessary and sufficient condition for OLS to be equivalent to GLS, is presented here, see Baltagi (1988). In Chapter 9, we saw that GLS is equivalent to OLS, for every y, if and only if

$$X'\Omega^{-1}\bar{P}_X = 0 \tag{10.9}$$

In this case, $X = \text{diag}[X_i]$, $\Omega^{-1} = \Sigma^{-1} \otimes I_T$, and $\bar{P}_X = \text{diag}[\bar{P}_{X_i}]$. Hence, the typical element of (10.9), see problem 1, is

$$\sigma^{ij}X_i'\bar{P}_{X_j} = 0 \tag{10.10}$$

This is automatically satisfied for $i = j$. For $i \neq j$, this holds if $\sigma^{ij} = 0$ or $X_i'\bar{P}_{X_j} = 0$. Note that $\sigma^{ij} = 0$ is the first sufficient condition provided by Zellner (1962). The latter condition $X_i'\bar{P}_{X_j} = 0$ implies that the set of regressors in the i-th equation are a perfect linear combination

of those in the j-th equation. Since $X_j'\bar{P}_{X_i} = 0$ has to hold also, X_j has to be a perfect linear combination of the regressors in the i-th equation. X_i and X_j span the same space. Both X_i and X_j have full column rank for OLS to be feasible, hence they have to be of the same dimension for $X_i'\bar{P}_{X_j} = X_j'\bar{P}_{X_i} = 0$. In this case, $X_i' = CX_j'$, where C is a nonsingular matrix, i.e., the regressors in the i-th equation are a perfect linear combination of those in the j-th equation. This includes the second sufficient condition derived by Zellner (1962). In practice, different economic behavioral equations contain different number of right hand side variables. In this case, one rearranges the SUR into blocks where each block has the same number of right hand side variables. For two equations (i and j) belonging to two different blocks ($i \neq j$), (10.10) is satisfied if the corresponding σ^{ij} is zero, i.e., Σ has to be block diagonal. However, in this case, GLS performed on the whole system is equivalent to GLS performed on each block taken separately. Hence, (10.10) is satisfied for SUR if it is satisfied for each block taken separately.

Revankar (1974) considered the case where X_2 is a subset of X_1. In this case, there is no gain in using SUR for estimating β_2. In fact, problem 2 asks the reader to verify that $\widehat{\beta}_{2,SUR} = \widehat{\beta}_{2,OLS}$. However, this is not the case for β_1. It is easy to show that $\widehat{\beta}_{1,SUR} = \widehat{\beta}_{1,OLS} - Ae_{2,OLS}$, where A is a matrix defined in problem 2, and $e_{2,OLS}$ are the OLS residuals for the second equation.

Telser (1964) suggested an iterative least squares procedure for SUR equations. For the two equations model given in (10.1), this estimation method involves the following:

1. Compute the OLS residuals e_1 and e_2 from both equations.
2. Include e_1 as an extra regressor in the second equation and e_2 as an extra regressor in the first equation. Compute the new least squares residuals and iterate this step until convergence of the estimated coefficients. The resulting estimator has the same asymptotic distribution as Zellner's (1962) SUR estimator.

Conniffe (1982) suggests stopping at the second step because in small samples this provides most of the improvement in precision. In fact, Conniffe (1982) argues that it may be unnecessary and even disadvantageous to calculate Zellner's estimator proper. Extensions to multiple equations is simple. Step 1 is the same where one computes least squares residuals of every equation. Step 2 adds the residuals of all other equations in the equation of interest. OLS is run and the new residuals are computed. One can stop at this second step or iterate until convergence.

10.2 Feasible GLS Estimation

In practice, Σ is not known and has to be estimated. Zellner (1962) recommended the following feasible GLS estimation procedure:

$$s_{ii} = \sum_{t=1}^{T} e_{it}^2/(T - K_i) \qquad \text{for} \qquad i = 1, 2 \tag{10.11}$$

and

$$s_{ij} = \sum_{t=1}^{T} e_{it}e_{jt}/(T - K_i)^{1/2}(T - K_j)^{1/2} \quad \text{for} \quad i, j = 1, 2 \quad \text{and} \quad i \neq j \tag{10.12}$$

where e_{it} denotes OLS residuals of the i-th equation. s_{ii} is the s^2 of the regression for the i-th equation. This is unbiased for σ_{ii}. However, s_{ij} for $i \neq j$ is not unbiased for σ_{ij}. In fact, the unbiased estimate is

$$\widetilde{s}_{ij} = \sum_{t=1}^{T} e_{it}e_{jt}/[T - K_i - K_j + \text{ tr } (B)] \qquad \text{for} \qquad i, j = 1, 2 \tag{10.13}$$

where $B = X_i(X_i'X_i)^{-1}X_i'X_j(X_j'X_j)^{-1}X_j' = P_{X_i}P_{X_j}$, see problem 4. Using this last estimator may lead to a variance-covariance matrix that is not positive definite. For consistency, however, all we need is a division by T, however this leaves us with a biased estimator:

$$\widehat{s}_{ij} = \sum_{t=1}^{T} e_{it}e_{jt}/T \quad \text{for} \quad i,j = 1,2 \tag{10.14}$$

Using this consistent estimator of Σ will result in feasible GLS estimates that are asymptotically efficient. In fact, if one iterates this procedure, i.e., compute feasible GLS residuals and second round estimates of Σ using these GLS residuals in (10.14), and continue iterating, until convergence, this will lead to maximum likelihood estimates of the regression coefficients, see Oberhofer and Kmenta (1974).

Relative Efficiency of OLS in the Case of Simple Regressions

To illustrate the gain in efficiency of Zellner's SUR compared to performing OLS on each equation separately, Kmenta(1986, pp.641-643) considers the following two simple regression equations:

$$\begin{aligned} Y_{1t} &= \beta_{11} + \beta_{12}X_{1t} + u_{1t} \\ Y_{2t} &= \beta_{21} + \beta_{22}X_{2t} + u_{2t} \quad \text{for} \quad t = 1,2,\ldots T; \end{aligned} \tag{10.15}$$

and proves that

$$\text{var}(\widehat{\beta}_{12,GLS})/\text{var}(\widehat{\beta}_{12,OLS}) = (1-\rho^2)/[1-\rho^2 r^2] \tag{10.16}$$

where ρ is the correlation coefficient between u_1 and u_2, and r is the sample correlation coefficient between X_1 and X_2. Problem 5 asks the reader to verify (10.16). In fact, the same relative efficiency ratio holds for β_{22}, i.e., $\text{var}(\widehat{\beta}_{22,GLS})/\text{var}(\widehat{\beta}_{22,OLS})$ is given by that in (10.16). This confirms the two results obtained above, namely, that as ρ increases this relative efficiency ratio decreases and OLS is less efficient than GLS. Also, as r increases this relative efficiency ratio increases and there is less gain in performing GLS rather than OLS. For $\rho = 0$ or $r = 1$, the efficiency ratio is 1, and OLS is equivalent to GLS. However, if ρ is large, say 0.9 and r is small, say 0.1 then (10.16) gives a relative efficiency of 0.11. For a tabulation of (10.16) for various values of ρ^2 and r^2, see Table 12-1 of Kmenta (1986, p. 642).

Relative Efficiency of OLS in the Case of Multiple Regressions

With more regressors in each equation, the relative efficiency story has to be modified, as indicated by Binkley and Nelson (1988). In the two equation model considered in (10.2) with K_1 regressors X_1 in the first equation and K_2 regressors X_2 in the second equation

$$\text{var}(\widehat{\beta}_{GLS}) = (X'\Omega^{-1}X)^{-1} = \begin{bmatrix} \sigma^{11}X_1'X_1 & \sigma^{12}X_1'X_2 \\ \sigma^{21}X_2'X_1 & \sigma^{22}X_2'X_2 \end{bmatrix}^{-1} = [A_{ij}] \tag{10.17}$$

If we focus on the regression estimates of the first equation, we get $\text{var}(\widehat{\beta}_{1,GLS}) = A_{11} = [\sigma^{11}X_1'X_1 - \sigma^{12}X_1'X_2(\sigma^{22}X_2'X_2)^{-1}\sigma^{21}X_2'X_1]^{-1}$ see problem 6. Using the fact that

$$\Sigma^{-1} = [1/(1-\rho^2)] \begin{bmatrix} 1/\sigma_{11} & -\rho^2/\sigma_{12} \\ -\rho^2/\sigma_{21} & 1/\sigma_{22} \end{bmatrix}$$

where $\rho^2 = \sigma_{12}^2/\sigma_{11}\sigma_{22}$, one gets

$$\text{var}(\widehat{\beta}_{1,GLS}) = [\sigma_{11}(1-\rho^2)]\{X_1'X_1 - \rho^2(X_1'P_{X_2}X_1)\}^{-1} \tag{10.18}$$

Add and subtract $\rho^2 X_1'X_1$ from the expression to be inverted, one gets

$$\text{var}(\widehat{\beta}_{1,GLS}) = \sigma_{11}\{X_1'X_1 + [\rho^2/(1-\rho^2)]E'E\}^{-1} \tag{10.19}$$

where $E = \bar{P}_{X_2}X_1$ is the matrix whose columns are the OLS residuals of each variable in X_1 regressed on X_2. If $E = 0$, there is no gain in SUR over OLS for the estimation of β_1. $X_1 = X_2$ or X_1 is a subset of X_2 are two such cases. One can easily verify that (10.19) is the variance-covariance matrix of an OLS regression with regressor matrix

$$W = \begin{bmatrix} X_1 \\ \theta E \end{bmatrix}$$

where $\theta^2 = \rho^2/(1-\rho^2)$. Now let us focus on the efficiency of the estimated coefficient of the q-th variable, X_q in X_1. Recall, from Chapter 4, that for the regression of y on X_1

$$\text{var}(\widehat{\beta}_{q,OLS}) = \sigma_{11}/\left\{\textstyle\sum_{t=1}^{T} x_{tq}^2(1-R_q^2)\right\} \tag{10.20}$$

where the denominator is the residual sum of squares of X_q on the other (K_1-1) regressors in X_1 and R_q^2 is the corresponding R^2 of that regression. Similarly, from (10.19),

$$\text{var}(\widehat{\beta}_{q,SUR}) = \sigma_{11}/\left\{\textstyle\sum_{t=1}^{T} x_{tq}^2 + \theta^2 \sum_{t=1}^{T} e_{tq}^2\right\}(1-R_q^{*2}) \tag{10.21}$$

where the denominator is the residual sum of squares of $\begin{bmatrix} X_q \\ \theta e_q \end{bmatrix}$ on the other (K_1-1) regressors in W, and R_q^{*2} is the corresponding R^2 of that regression. Add and subtract $\sum_{t=1}^{T} x_{tq}^2(1-R_q^2)$ to the denominator of (10.21), we get

$$\text{var}(\widehat{\beta}_{q,SUR}) = \frac{\sigma_{11}}{\left\{\sum_{t=1}^{T} x_{tq}^2(1-R_q^2) + \sum_{t=1}^{T} x_{tq}^2(R_q^2-R_q^{*2}) + \theta^2 \sum_{t=1}^{T} e_{tq}^2(1-R_q^{*2})\right\}} \tag{10.22}$$

This variance differs from $\text{var}(\widehat{\beta}_{q,OLS})$ in (10.20) by the two extra terms in the denominator. If $\rho = 0$, then $\theta^2 = 0$, so that $W' = [X_1', 0]$ and $R_q^2 = R_q^{*2}$. In this case, (10.22) reduces to (10.20). If X_q also appears in the second equation, or in general is spanned by the variables in X_2, then $e_{tq} = 0$, $\sum_{t=1}^{T} e_{tq}^2 = 0$ and from (10.22) there is gain in efficiency only if $R_q^2 \geq R_q^{*2}$. R_q^2 is a measure of multicollinearity of X_q with the other (K_1-1) regressors in the first equation, i.e., X_1. If this is high, then it is more likely for $R_q^2 \geq R_q^{*2}$. Therefore, the higher the multicollinearity within X_1, the greater the potential for a decrease in variance of OLS by SUR. Note that $R_q^2 = R_q^{*2}$ when $\theta E = 0$. This is true if $\theta = 0$, or $E = 0$. The latter occurs when X_1 is spanned by the sub-space of X_2. Problem 7 asks the reader to verify that $R_q^2 = R_q^{*2}$ when X_1 is orthogonal to X_2. Therefore, with more regressors in each equation, one has to consider the correlation between the X's within each equation as well as that across equations. Even when the X's *across* equations are highly correlated, there may still be gains from joint estimation using SUR when there is high mulicollinearity *within* each equation.

10.3 Testing Diagonality of the Variance-Covariance Matrix

Since the diagonality of Σ is at the heart of using SUR estimation methods, it is important to look at tests for H_0: Σ is diagonal. Breusch and Pagan (1980) derived a simple and easy to use Lagrange multiplier statistic for testing H_0. This is based upon the sample correlation coefficients of the OLS residuals:

$$LM = T \sum_{i=2}^{M} \sum_{j=1}^{i-1} r_{ij}^2 \tag{10.23}$$

where M denotes the number of equations and $r_{ij} = \widehat{s}_{ij}/(\widehat{s}_{ii}\widehat{s}_{jj})^{1/2}$. The $\widehat{s}_{ij}$'s are computed from OLS residuals as in (10.14). Under the null hypothesis, λ_{LM} has an asymptotic $\chi^2_{M(M-1)/2}$ distribution. Note that the $\widehat{s}_{ij}$'s are needed for feasible GLS estimation. Therefore, it is easy to compute the r_{ij}'s and λ_{LM} by summing the squares of half the number of off-diagonal elements of $R = [r_{ij}]$ and multiplying the sum by T. For example, for the two equations case, $\lambda_{LM} = Tr_{21}^2$ which is asymptotically distributed as χ_1^2 under H_0. For the three equations case, $\lambda_{LM} = T(r_{21}^2 + r_{31}^2 + r_{32}^2)$ which is asymptotically distributed as χ_3^2 under H_0.

Alternatively, the Likelihood Ratio test can also be used to test for diagonality of Σ. This is based on the determinants of the variance covariance matrices estimated by MLE for the restricted and unrestricted models:

$$\lambda_{LR} = T\left(\sum_{i=1}^{M} \log \widehat{s}_{ii} - \log|\widehat{\Sigma}|\right) \tag{10.24}$$

where $\widehat{s}_{ii}$ is the restricted MLE of σ_{ii} obtained from the OLS residuals as in (10.14). The matrix $\widehat{\Sigma}$ denotes the unrestricted MLE of Σ. This may be adequately approximated with an estimator based on the feasible GLS estimator $\widehat{\beta}_{FGLS}$, see Judge et al. (1982). Under H_0, λ_{LR} has an asymptotic $\chi^2_{M(M-1)/2}$ distribution.

10.4 Seemingly Unrelated Regressions with Unequal Observations

Srivastava and Dwivedi (1979) surveyed the developments in the SUR model and described the extensions of this model to the serially correlated case, the nonlinear case, the misspecified case, and that with unequal number of observations. Srivastava and Giles (1988) dedicated a monograph to SUR models, and surveyed the finite sample as well as asymptotic results. More recently, Fiebig (2001) gives a concise and up to date account of research in this area. In this section, we consider one extension to focus upon. This is the case of SUR with unequal number of observations considered by Schmidt (1977), Baltagi, Garvin and Kerman (1989) and Hwang (1990).

Let the first firm have T observations common with the second firm, but allow the latter to have N extra observations. In this case, (10.2) will have y_1 of dimension $T \times 1$ whereas y_2 will be of dimension $(T + N) \times 1$. In fact, $y_2' = (y_2^{*\prime}, y_2^{o\prime})$ and $X_2' = (X_2^{*\prime}, X_2^{o\prime})$ with $*$ denoting the T common observations for the second firm, and o denoting the extra N observations for the second firm. The disturbances will now have a variance-covariance matrix

$$\Omega = \begin{bmatrix} \sigma_{11} I_T & \sigma_{12} I_T & 0 \\ \sigma_{12} I_T & \sigma_{22} I_T & 0 \\ 0 & 0 & \sigma_{22} I_N \end{bmatrix} \tag{10.25}$$

GLS on (10.2) will give

$$\widehat{\beta}_{GLS} = \begin{bmatrix} \sigma^{11}X_1'X_1 & \sigma^{12}X_1'X_2^* \\ \sigma^{12}X_2^{*\prime}X_1 & \sigma^{22}X_2^{*\prime}X_2^* + (X_2^{o\prime}X_2^o)/\sigma_{22}) \end{bmatrix}^{-1} \begin{bmatrix} \sigma^{11}X_1'y_1 + \sigma^{12}X_1'y_2^* \\ \sigma^{12}X_2^{*\prime}y_1 + \sigma^{22}X_2^{*\prime}y_2^* + (X_2^{o\prime}y_2^o)/\sigma_{22}) \end{bmatrix} \tag{10.26}$$

where $\Sigma^{-1} = [\sigma^{ij}]$ for $i, j = 1, 2$. If we run OLS on each equation (T for the first equation, and $T + N$ for the second equation) and denote the residuals for the two equations by e_1 and e_2, respectively, then we can partition the latter residuals into $e_2' = (e_2^{*\prime}, e_2^{o\prime})$. In order to estimate Ω, Schmidt (1977) considers the following procedures:

(1) Ignore the extra N observations in estimating Ω. In this case

$$\widehat{\sigma}_{11} = s_{11} = e_1'e_1/T;\ \widehat{\sigma}_{12} = s_{12} = e_1'e_2^*/T \quad \text{and} \quad \widehat{\sigma}_{22} = s_{22}^* = e_2^{*\prime}e_2^*/T \tag{10.27}$$

(2) Use $T + N$ observations to estimate σ_{22}. In other words, use s_{11}, s_{12} and $\widehat{\sigma}_{22} = s_{22} = e_2'e_2/(T + N)$. This procedure is attributed to Wilks (1932) and has the disadvantage of giving estimates of Ω that are not positive definite.

(3) Use s_{11} and s_{22}, but modify the estimate of σ_{12} such that $\widehat{\Omega}$ is positive definite. Srivastava and Zaatar (1973) suggest $\widehat{\sigma}_{12} = s_{12}(s_{22}/s_{22}^*)^{1/2}$.

(4) Use all $(T + N)$ observations in estimating Ω. Hocking and Smith (1968) suggest using $\widehat{\sigma}_{11} = s_{11} - (N/N + T)(s_{12}/s_{22}^*)^2(s_{22}^* - s_{22}^o)$ where $s_{22}^o = e_2^{o\prime}e_2^o/N$; $\widehat{\sigma}_{12} = s_{12}(s_{22}/s_{22}^*)$ and $\widehat{\sigma}_{22} = s_{22}$.

(5) Use a maximum likelihood procedure.

All estimators of Ω are consistent, and $\widehat{\beta}_{FGLS}$ based on any of these estimators will be asymptotically efficient. Schmidt considers their small sample properties by means of Monte Carlo experiments. Using the set up of Kmenta and Gilbert (1968) he finds for $T = 10, 20, 50$ and $N = 5, 10, 20$ and various correlation of the X's and the disturbances across equations the following disconcerting result: "..it is certainly remarkable that procedures that essentially ignore the extra observations in estimating Σ (e.g., Procedure 1) do not generally do badly relative to procedures that use the extra observations fully (e.g., Procedure 4 or MLE). Except when the disturbances are highly correlated across equations, we may as well just forget about the extra observations in estimating Σ. This is not an intuitively reasonable procedure."

Hwang (1990) re-parametrizes these estimators in terms of the elements of Σ^{-1} rather than Σ. After all, it is Σ^{-1} rather than Σ that appears in the GLS estimator of β. This re-parametrization shows that the estimators of Σ^{-1} no longer have the ordering in terms of their use of the extra observations as that reported by Schmidt (1977). However, regardless of the parametrization chosen, it is important to point out that *all* the observations are used in the estimation of β whether at the first stage for obtaining the least squares residuals, or in the final stage in computing GLS. Baltagi et al. (1989) show using Monte Carlo experiments that better estimates of Σ or its inverse Σ^{-1} in Mean Square Error sense, do not necessarily lead to better GLS estimates of β.

10.5 Empirical Example

Baltagi and Griffin (1983) considered the following gasoline demand equation:

$$\log\frac{Gas}{Car} = \alpha + \beta_1 \log\frac{Y}{N} + \beta_2 \log\frac{P_{MG}}{P_{GDP}} + \beta_3 \log\frac{Car}{N} + u$$

where Gas/Car is motor gasoline consumption per auto, Y/N is real per capita income, P_{MG}/P_{GDP} is real motor gasoline price and Car/N denotes the stock of cars per capita. This data consists of annual observations across 18 OECD counties, covering the period 1960-1978. It is provided as GASOLINE.DAT on the Springer web site. We consider the first two countries: Austria and Belgium. OLS on this data yields

$$\begin{array}{lcccccccc} \text{Austria} & \log\frac{Gas}{Car} & = & 3.727 & + & 0.761 \log\frac{Y}{N} & - & 0.793 \log\frac{P_{MG}}{P_{GDP}} & - & 0.520 \log\frac{Car}{N} \\ & & & (0.373) & & (0.211) & & (0.150) & & (0.113) \\ \text{Belgium} & \log\frac{Gas}{Car} & = & 3.042 & + & 0.845 \log\frac{Y}{N} & - & 0.042 \log\frac{P_{MG}}{P_{GDP}} & - & 0.673 \log\frac{Car}{N} \\ & & & (0.453) & & (0.170) & & (0.158) & & (0.093) \end{array}$$

where the standard errors are shown in parentheses. Based on these OLS residuals, the estimate of Σ is given by

$$\widehat{\Sigma} = \begin{bmatrix} 0.0012128 & 0.00023625 \\ & 0.00092367 \end{bmatrix}$$

The Seemingly Unrelated Regression estimates based on this $\widehat{\Sigma}$, i.e., after one iteration, are given by

$$\begin{array}{lcccccccc} \text{Austria} & \log\frac{Gas}{Car} & = & 3.713 & + & 0.721 \log\frac{Y}{N} & - & 0.754 \log\frac{P_{MG}}{P_{GDP}} & - & 0.496 \log\frac{Car}{N} \\ & & & (0.372) & & (0.209) & & (0.146) & & (0.111) \\ \text{Belgium} & \log\frac{Gas}{Car} & = & 2.843 & + & 0.835 \log\frac{Y}{N} & - & 0.131 \log\frac{P_{MG}}{P_{GDP}} & - & 0.686 \log\frac{Car}{N} \\ & & & (0.445) & & (0.170) & & (0.154) & & (0.093) \end{array}$$

The Breusch-Pagan (1980) Lagrange multiplier test for diagonality of Σ is $Tr_{21}^2 = 0.947$ which is distributed as χ_1^2 under the null hypothesis. The Likelihood Ratio test for the diagonality of Σ, given in (10.23), yields a value of 1.778 which is also distributed as χ_1^2 under the null hypothesis. Both test statistics do not reject H_0. These SUR results were run using SHAZAM and could be iterated further. Note the reduction in the standard errors of the estimated regression coefficients is minor as we compare the OLS and SUR estimates.

Suppose that we only have the first 15 observations (1960-1974) on Austria and all 19 observations (1960-1978) on Belgium. We now apply the four feasible GLS procedures described by Schmidt (1977). The first procedure which ignores the extra 4 observations in estimating Σ yields $s_{11} = 0.00086791$, $s_{12} = 0.00026357$ and $s_{22}^* = 0.00109947$ as described in (10.27). The

resulting SUR estimates are given by

$$\text{Austria} \quad \log\frac{Gas}{Car} = \underset{(0.438)}{4.484} + \underset{(0.168)}{0.817}\log\frac{Y}{N} - \underset{(0.176)}{0.580}\log\frac{P_{MG}}{P_{GDP}} - \underset{(0.098)}{0.487}\log\frac{Car}{N}$$

$$\text{Belgium} \quad \log\frac{Gas}{Car} = \underset{(0.436)}{2.936} + \underset{(0.164)}{0.848}\log\frac{Y}{N} - \underset{(0.151)}{0.095}\log\frac{P_{MG}}{P_{GDP}} - \underset{(0.090)}{0.686}\log\frac{Car}{N}$$

The second procedure, due to Wilks (1932) uses the same s_{11} and s_{12} in procedure 1, but $\widehat{\sigma}_{22} = s_{22} = e_2'e_2/19 = 0.00092367$. The resulting SUR estimates are given by

$$\text{Austria} \quad \log\frac{Gas}{Car} = \underset{(0.437)}{4.521} + \underset{(0.167)}{0.806}\log\frac{Y}{N} - \underset{(0.174)}{0.554}\log\frac{P_{MG}}{P_{GDP}} - \underset{(0.098)}{0.476}\log\frac{Car}{N}$$

$$\text{Belgium} \quad \log\frac{Gas}{Car} = \underset{(0.399)}{2.937} + \underset{(0.150)}{0.848}\log\frac{Y}{N} - \underset{(0.138)}{0.094}\log\frac{P_{MG}}{P_{GDP}} - \underset{(0.082)}{0.685}\log\frac{Car}{N}$$

The third procedure based on Srivastava and Zaatar (1973) use the same s_{11} and s_{22} as procedure 2, but modify $\widehat{\sigma}_{12} = s_{12}(s_{22}/s^*_{22})^{1/2} = 0.00024158$. The resulting SUR estimates are given by

$$\text{Austria} \quad \log\frac{Gas}{Car} = \underset{(0.438)}{4.503} + \underset{(0.168)}{0.812}\log\frac{Y}{N} - \underset{(0.176)}{0.567}\log\frac{P_{MG}}{P_{GDP}} - \underset{(0.098)}{0.481}\log\frac{Car}{N}$$

$$\text{Belgium} \quad \log\frac{Gas}{Car} = \underset{(0.400)}{2.946} + \underset{(0.151)}{0.847}\log\frac{Y}{N} - \underset{(0.139)}{0.090}\log\frac{P_{MG}}{P_{GDP}} - \underset{(0.082)}{0.684}\log\frac{Car}{N}$$

The fourth procedure due to Hocking and Smith (1968) yields $\widehat{\sigma}_{11} = 0.00085780$, $\widehat{\sigma}_{12} = 0.0022143$ and $\widehat{\sigma}_{22} = s_{22} = 0.00092367$. The resulting SUR estimates are given by

$$\text{Austria} \quad \log\frac{Gas}{Car} = \underset{(0.437)}{4.485} + \underset{(0.168)}{0.817}\log\frac{Y}{N} - \underset{(0.176)}{0.579}\log\frac{P_{MG}}{P_{GDP}} - \underset{(0.098)}{0.487}\log\frac{Car}{N}$$

$$\text{Belgium} \quad \log\frac{Gas}{Car} = \underset{(0.400)}{2.952} + \underset{(0.151)}{0.847}\log\frac{Y}{N} - \underset{(0.139)}{0.086}\log\frac{P_{MG}}{P_{GDP}} - \underset{(0.082)}{0.684}\log\frac{Car}{N}$$

In this case, there is not much difference among these four alternative estimates.

Problems

1. (a) Show that OLS on a system of two Zellner's SUR equations given in (10.2) is the same as OLS on each equation taken separately. What about the estimated variance-covariance matrix of the coefficients? Will they be the same?

   (b) In the General Linear Model, we found a necessary and sufficient condition for OLS to be equivalent to GLS is that $X'\Omega^{-1}\bar{P}_X = 0$ for every y where $\bar{P}_X = I - P_X$. Show that a necessary and sufficient condition for Zellner's GLS to be equivalent to OLS is that $\sigma^{ij}X_i'\bar{P}_{X_j} = 0$ for $i \neq j$ as described in (10.10). This is based on Baltagi (1988).

   (c) Show that the two sufficient conditions given by Zellner for SUR to be equivalent to OLS both satisfy the necessary and sufficient condition given in part (b).

(d) Show that if $X_i = X_j C'$ where C is an arbitrary nonsingular matrix, then the necessary and sufficient condition given in part (b) is satisfied.

2. For the two SUR equations given in (10.2). Let $X_1 = (X_2, X_e)$, i.e., X_2 is a subset of X_1. Prove that

   (a) $\widehat{\beta}_{2,SUR} = \widehat{\beta}_{2,OLS}$.

   (b) $\widehat{\beta}_{1,SUR} = \widehat{\beta}_{1,OLS} - Ae_{2,OLS}$, where $A = \widehat{s}_{12}(X_1'X_1)^{-1}X_1'/\widehat{s}_{22}$. $e_{2,OLS}$ are the OLS residuals from the second equation, and the $\widehat{s}_{ij}$'s are defined in (10.14).

3. For the two SUR equations given in (10.2). Let X_1 and X_2 be orthogonal, i.e., $X_1'X_2 = 0$. Show that knowing the true Σ we get

   (a) $\widehat{\beta}_{1,GLS} = \widehat{\beta}_{1,OLS} + (\sigma^{12}/\sigma^{11})(X_1'X_1)^{-1}X_1'y_2$ and $\widehat{\beta}_{2,GLS} = \widehat{\beta}_{2,OLS} + (\sigma^{21}/\sigma^{22})(X_2'X_2)^{-1}X_2'y_1$.

   (b) What are the variances of these estimates?

   (c) If X_1 and X_2 are single regressors, what are the relative efficiencies of $\widehat{\beta}_{i,OLS}$ with respect to $\widehat{\beta}_{i,GLS}$ for $i = 1, 2$?

4. Verify that $\widetilde{s}_{ij}$, given in (10.13), is unbiased for σ_{ij}. Note that for computational purposes $\text{tr}(B) = \text{tr}(P_{X_i}P_{X_j})$.

5. *Relative Efficiency of OLS in the Case of Simple Regressions.* This is based on Kmenta (1986, pp. 641-643). For the system of two equations given in (10.15), show that

   (a) $\text{var}(\widehat{\beta}_{12,OLS}) = \sigma_{11}/m_{x_1x_1}$ and $\text{var}(\widehat{\beta}_{22,OLS}) = \sigma_{22}/m_{x_2x_2}$ where $m_{x_ix_j} = \sum_{t=1}^{T}(X_{it} - \bar{X}_i)(X_{jt} - \bar{X}_j)$ for $i, j = 1, 2$.

   (b) $$\text{var}\begin{pmatrix} \widehat{\beta}_{12,GLS} \\ \widehat{\beta}_{22,GLS} \end{pmatrix} = (\sigma_{11}\sigma_{22} - \sigma_{12}^2)\begin{bmatrix} \sigma_{22}m_{x_1x_1} & -\sigma_{12}m_{x_1x_2} \\ -\sigma_{12}m_{x_1x_2} & \sigma_{11}m_{x_2x_2} \end{bmatrix}^{-1}.$$
   Deduce that $\text{var}\left(\widehat{\beta}_{12,GLS}\right) = (\sigma_{11}\sigma_{22} - \sigma_{12}^2)\sigma_{11}m_{x_2x_2}/[\sigma_{11}\sigma_{22}m_{x_2x_2}m_{x_1x_1} - \sigma_{12}^2 m_{x_1x_2}^2]$ and $\text{var}\left(\widehat{\beta}_{22,GLS}\right) = (\sigma_{11}\sigma_{22} - \sigma_{12}^2)\sigma_{22}m_{x_1x_1}/[\sigma_{11}\sigma_{22}m_{x_1x_1}m_{x_2x_2} - \sigma_{12}^2 m_{x_1x_2}^2]$.

   (c) Using $\rho = \sigma_{12}/(\sigma_{11}\sigma_{22})^{1/2}$ and $r = m_{x_1x_2}/(m_{x_1x_1}m_{x_2x_2})^{1/2}$ and the results in parts (a) and (b), show that (10.16) holds, i.e., $\text{var}(\widehat{\beta}_{12,GLS})/\text{var}(\widehat{\beta}_{12,OLS}) = (1 - \rho^2)/[1 - \rho^2 r^2]$.

   (d) Differentiate (10.16) with respect to $\theta = \rho^2$ and show that (10.16) is a non-increasing function of θ. Similarly, differentiate (10.16) with respect to $\lambda = r^2$ and show that (10.16) is a non-decreasing function of λ. Finally, compute this efficiency measure (10.16) for various values of ρ^2 and r^2 between 0 and 1 at 0.1 intervals, see Kmenta's (1986) Table 12-1, p.642.

6. *Relative Efficiency of OLS in the Case of Multiple Regressions.* This is based on Binkley and Nelson (1988). Using partitioned inverse formulas, verify that $\text{var}(\widehat{\beta}_{1,GLS}) = A_{11}$ given below (10.17). Deduce (10.18) and (10.19).

7. Consider the multiple regression case with *orthogonal* regressors across the two equations, i.e., $X_1'X_2 = 0$. Verify that $R_q^2 = R_q^{*2}$, where R_q^2 and R_q^{*2} are defined below (10.20) and (10.21), respectively.

8. (a) *SUR With Unequal Number of Observations.* This is based on Schmidt (1977). Derive the GLS estimator for SUR with unequal number of observations given in (10.26).

   (b) Show that if $\sigma_{12} = 0$, SUR with unequal number of observations reduces to OLS on each equation separately.

9. Grunfeld (1958) considered the following investment equation:

$$I_{it} = \alpha + \beta_1 F_{it} + \beta_2 C_{it} + u_{it}$$

where I_{it} denotes real gross investment for firm i in year t, F_{it} is the real value of the firm (shares outstanding) and C_{it} is the real value of the capital stock. This data set consists of 10 large U.S. manufacturing firms over 20 years, 1935-1954, and are given in Boot and de Witt (1960). It is provided as GRUNFELD.DAT on the Springer web site. Consider the first three firms: G.M., U.S. Steel and General Electric.

   (a) Run OLS of I on a constant, F and C for each of the 3 firms separately. Plot the residuals against time. Print the variance-covariance of the estimates.

   (b) Test for serial correlation in each regression.

   (c) Run Seemingly Unrelated Regressions (SUR) for the first two firms. Compare with OLS.

   (d) Run SUR for the three assigned firms. Compare these results with those in part (c).

   (e) Test for the diagonality of Σ across these three equations.

   (f) Test for the equality of all coefficients across the 3 firms.

10. *(Continue problem 9).* Consider the first two firms again and focus on the coefficient of F. Refer to the Binkley and Nelson (1988) article in *The American Statistician*, and compute R_q^2, R_q^{*2}, Σe_{tq}^2 and Σx_{tq}^2.

   (a) What would be equations (10.20) and (10.21) for your data set?

   (b) Substitute estimates of σ_{11} and θ^2 and verify that the results are the same as those obtained in problems 9(a) and 9(c).

   (c) Compare the results from equations (10.20) and (10.21) in part (a). What do you conclude?

11. *(Continue problem 9).* Consider the first two firms once more. Now you only have the first 15 observations on the first firm and all 20 observations on the second firm. Apply Schmidt's (1977) feasible GLS estimators and compare the resulting estimates.

12. For the Baltagi and Griffin (1983) Gasoline Data considered in section 10.5, the model is

$$\log\frac{Gas}{Car} = \alpha + \beta_1 \log\frac{Y}{N} + \beta_2 \log\frac{P_{MG}}{P_{GDP}} + \beta_3 \log\frac{Car}{N} + u$$

where Gas/Car is motor gasoline consumption per auto, Y/N is real per capita income, P_{MG}/P_{GDP} is real motor gasoline price and Car/N denotes the stock of cars per capita.

   (a) Run Seemingly Unrelated Regressions (SUR) for the first two countries. Compare with OLS.

   (b) Run SUR for the first three countries. Comment on the results and compare with those of part (a). (Are there gains in efficiency?)

   (c) Test for Diagonality of Σ across the three equations using the Breusch and Pagan (1980) LM test and the Likelihood Ratio test.

   (d) Test for the equality of all coefficients across the 3 countries.

   (e) Consider the first 2 countries once more. Now you only have the first 15 observations on the first country and all 19 observations on the second country. Apply Schmidt's (1977) feasible GLS estimators, and compare the results.

13. *Trace Minimization of Singular Systems with Cross-Equation Restrictions.* This is based on Baltagi (1993). Berndt and Savin (1975) demonstrated that when certain cross-equation restrictions are imposed, restricted least squares estimation of a singular set of SUR equations will not be invariant to which equation is deleted. Consider the following set of three equations with the same regressors:

$$y_i = \alpha_i \iota_T + \beta_i X + \epsilon_i \quad i = 1, 2, 3.$$

where $y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$, $X = (x_1, x_2, \ldots, x_T)'$, and ϵ_i for $(i = 1, 2, 3)$ are $T \times 1$ vectors and ι_T is a vector of ones of dimension T. α_i and β_i are scalars, and these equations satisfy the adding up restriction $\sum_{i=1}^{3} y_{it} = 1$ for every $t = 1, 2, \ldots, T$. Additionally, we have a cross-equation restriction: $\beta_1 = \beta_2$.

(a) Denote the unrestricted OLS estimates of β_i by b_i where $b_i = \sum_{t=1}^{T}(x_t - \bar{x})y_{it} / \sum_{t=1}^{T}(x_t - \bar{x})^2$ for $i = 1, 2, 3$, and $\bar{x} = \sum_{t=1}^{T} x_t / T$. Show that these unrestricted b_i's satisfy the adding up restriction $\beta_1 + \beta_2 + \beta_3 = 0$ on the true parameters automatically.

(b) Show that if one drops the first equation for $i = 1$ and estimate the remaining system by trace minimization subject to $\beta_1 = \beta_2$, one gets $\widehat{\beta}_1 = 0.4b_1 + 0.6b_2$.

(c) Now drop the second equation for $i = 2$, and show that estimating the remaining system by trace minimization subject to $\beta_1 = \beta_2$, gives $\widehat{\beta}_1 = 0.6b_1 + 0.4b_2$.

(d) Finally, drop the third equation for $i = 3$, and show that estimating the remaining system by trace minimization subject to $\beta_1 = \beta_2$ gives $\widehat{\beta}_1 = 0.5b_1 + 0.5b_2$.

Note that this also means the variance of $\widehat{\beta}_1$ is not invariant to the deleted equation. Also, this non-invariancy affects Zellner's SUR estimation if the restricted least squares residuals are used rather than the unrestricted least squares residuals in estimating the variance covariance matrix of the disturbances. **Hint:** See the solution by Im (1994).

14. For the Natural Gas data considered in Chapter 4, problem 16. The model is

$$\begin{aligned} \log Cons_{it} &= \beta_0 + \beta_1 \log Pg_{it} + \beta_2 \log Po_{it} + \beta_3 \log Pe_{it} + \beta_4 \log HDD_{it} \\ &\quad + \beta_5 \log PI_{it} + u_{it} \end{aligned}$$

where $i = 1, 2, \ldots, 6$ states and $t = 1, 2, \ldots, 23$ years.

(a) Run Seemingly Unrelated Regressions (SUR) for the first two states. Compare with OLS.

(b) Run SUR for all six states. Comment on the results and compare with those of part (a). (Are there gains in efficiency?)

(c) Test for Diagonality of Σ across the six states using the Breusch and Pagan (1980) LM test and the Likelihood Ratio test.

(d) Test for the equality of all coefficients across the six states.

15. *Equivalence of LR Test and Hausman Test.* This is based on Qian (1998). Suppose that we have the following two equations:

$$y_{gt} = \alpha_g + u_{gt} \qquad g = 1, 2, \qquad t = 1, 2, \ldots T$$

where (u_{1t}, u_{2t}) is normally distributed with mean zero and variance $\Omega = \Sigma \otimes I_T$ where $\Sigma = [\sigma_{gs}]$ for $g, s = 1, 2$. This is a simple example of the same regressors across two equations.

(a) Show that the OLS estimator of α_g is the same as the GLS estimator of α_g and both are equal to $\bar{y}_g = \sum_{t=1}^{T} y_{gt}/T$ for $g = 1, 2$.

(b) Derive the maximum likelihood estimators of α_g and σ_{gs} for $g, s, = 1, 2$. Compute the log-likelihood function evaluated at these unrestricted estimates.

(c) Compute the maximum likelihood estimators of α_g and σ_{gs} for $g, s = 1, 2$ under the null hypothesis H_0; $\sigma_{11} = \sigma_{22}$.

(d) Using parts (b) and (c) compute the LR test for H_0; $\sigma_{11} = \sigma_{22}$.

(e) Show that the LR test for H_0 derived in part (c) is asymptotically equivalent to the Hausman test based on the difference in estimators obtained in parts (b) and (c). Hausman's test is studied in Chapter 12.

16. *Estimation of a Triangular, Seemingly Unrelated Regression System by OLS.* This is based on Sentana (1997). Consider a system of three SUR equations in which the explanatory variables for the first equation are a subset of the explanatory variables for the second equation, which are in turn a subset of the explanatory variables for the third equation.

(a) Show that SUR applied to the first two equations is the same (for those equations) as SUR applied to all three equations. **Hint:** See Schmidt (1978).

(b) Using part (a) show that SUR for the first equation is equivalent to OLS.

(c) Using parts (a) and (b) show that SUR for the second equation is equivalent to OLS on the second equation with one additional regressor. The extra regressor is the OLS residuals from the first equation. **Hint:** Use Telser's (1964) results.

(d) Using parts (a), (b) and (c) show that SUR for the third equation is equivalent to OLS on the third equation with the residuals from the regressions in parts (b) and (c) as extra regressors.

References

This chapter is based on Zellner(1962), Kmenta(1986), Baltagi (1988), Binkley and Nelson (1988), Schmidt (1977) and Judge et al. (1982). References cited are:

Baltagi, B.H. (1988), "The Efficiency of OLS in a Seemingly Unrelated Regressions Model," *Econometric Theory*, Problem 88.3.4, 4: 536-537.

Baltagi, B.H. (1993), "Trace Minimization of Singular Systems With Cross-Equation Restrictions," *Econometric Theory*, Problem 93.2.4, 9: 314-315.

Baltagi, B., S. Garvin and S. Kerman (1989), "Further Monte Carlo Evidence on Seemingly Unrelated Regressions with Unequal Number of Observations," *Annales D'Economie et de Statistique*, 14: 103-115.

Baltagi, B.H. and J.M. Griffin (1983), "Gasoline Demand in the OECD: An Application of Pooling and Testing Procedures," *European Economic Review*, 22: 117-137.

Berndt, E.R. (1991), *The Practice of Econometrics: Classic and Contemporary* (Addison- Wesley: Reading, MA).

Berndt, E.R. and N.E. Savin (1975), "Estimation and Hypothesis Testing in Singular Equation Systems With Autoregressive Disturbances," *Econometrica*, 43: 937-957.

Binkley, J.K. and C.H. Nelson (1988), "A Note on the Efficiency of Seemingly Unrelated Regression," *The American Statistician*, 42: 137-139.

Boot, J. and G. de Witt (1960), "Investment Demand: An Empirical Contribution to the Aggregation Problem," *International Economic Review*, 1: 3-30.

Breusch, T.S. and A.R. Pagan (1980), "The Lagrange Multiplier Test and Its Applications to Model Specification in Econometrics," *Review of Economic Studies*, 47: 239-253.

Conniffe, D. (1982), "A Note on Seemingly Unrelated Regressions," *Econometrica*, 50: 229-233.

Dwivedi, T.D. and V.K. Srivastava (1978), "Optimality of Least Squares in the Seemingly Unrelated Regression Equations Model," *Journal of Econometrics*, 7: 391-395.

Fiebig, D.G. (2001), "Seemingly Unrelated Regression," Chapter 5 in Baltagi, B.H. (ed.), *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Grunfeld, Y. (1958), "The Determinants of Corporate Investment," unpublished Ph.D. dissertation (University of Chicago: Chicago, IL).

Hocking, R.R. and W.B. Smith (1968), "Estimation of Parameters in the Multivariate Normal Distribution with Missing Observations," *Journal of the American Statistical Association*, 63: 154-173.

Hwang, H.S. (1990), "Estimation of Linear SUR Model With Unequal Numbers of Observations," *Review of Economics and Statistics*, 72: 510-515.

Im, Eric Iksoon (1994), "Trace Minimization of Singular Systems With Cross-Equation Restrictions," *Econometric Theory*, Solution 93.2.4, 10: 450.

Kmenta, J. and R. Gilbert (1968), "Small Sample Properties of Alternative Estimators of Seemingly Unrelated Regressions," *Journal of the American Statistical Association*, 63: 1180-1200.

Milliken, G.A. and M. Albohali (1984), "On Necessary and Sufficient Conditions for Ordinary Least Squares Estimators to be Best Linear Unbiased Estimators," *The American Statistician*, 38: 298-299.

Oberhofer, W. and J. Kmenta (1974), "A General Procedure for Obtaining Maximum Likelihood Estimates in Generalized Regression Models," *Econometrica*, 42: 579-590.

Qian, H. (1998), "Equivalence of LR Test and Hausman Test," *Econometric Theory*, Problem 98.1.3, 14: 151.

Revankar, N.S. (1974), "Some Finite Sample Results in the Context of Two Seemingly Unrelated Regression Equations," *Journal of the American Statistical Association*, 71: 183-188.

Rossi, P.E. (1989), "The ET Interview: Professor Arnold Zellner," *Econometric Theory*, 5: 287-317.

Schmidt, P. (1977), "Estimation of Seemingly Unrelated Regressions With Unequal Numbers of Observations," *Journal of Econometrics*, 5: 365-377.

Schmidt, P. (1978), "A Note on the Estimation of Seemingly Unrelated Regression Systems," *Journal of Econometrics*, 7: 259-261.

Sentana, E. (1997), "Estimation of a Triangular, Seemingly Unrelated, Regression System by OLS," *Econometric Theory*, Problem 97.2.2, 13: 463.

Srivastava, V.K. and T.D. Dwivedi (1979), "Estimation of Seemingly Unrelated Regression Equations: A Brief Survey," *Journal of Econometrics*, 10: 15-32.

Srivastava, V.K. and D.E.A. Giles (1987), *Seemingly Unrelated Regression Equations Models: Estimation and Inference* (Marcel Dekker: New York).

Srivastava, J.N. and M.K. Zaatar (1973), "Monte Carlo Comparison of Four Estimators of Dispersion Matrix of a Bivariate Normal Population, Using Incomplete Data," *Journal of the American Statistical Association*, 68: 180-183.

Telser, L. (1964), "Iterative Estimation of a Set of Linear Regression Equations," *Journal of the American Statistical Association*, 59: 845-862.

Wilks, S.S. (1932), "Moments and Distributions of Estimates of Population Parameters From Fragmentary Samples," *Annals of Mathematical Statistics*, 3: 167-195.

Zellner, A. (1962), "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, 57: 348-368.

CHAPTER 11
Simultaneous Equations Model

11.1 Introduction

Economists formulate models for consumption, production, investment, money demand and money supply, labor demand and labor supply to attempt to explain the workings of the economy. These behavioral equations are estimated equation by equation or jointly as a system of equations. These are known as *simultaneous equations models*. Much of today's econometrics have been influenced and shaped by a group of economists and econometricians known as the Cowles Commission who worked together at the University of Chicago in the late 1940's, see Chapter 1. Simultaneous equations models had their genesis in economics during that period. Haavelmo's (1944) work emphasized the use of the probability approach to formulating econometric models. Koopmans and Marschak (1950) and Koopmans and Hood (1953) in two influential Cowles Commission monographs provided the appropriate statistical procedures for handling simultaneous equations models. In this chapter, we first give simple examples of simultaneous equations models and show why the least squares estimator is no longer appropriate. Next, we discuss the important problem of identification and give a simple necessary but not sufficient condition that helps check whether a specific equation is identified. Sections 11.2 and 11.3 give the estimation of a single and a system of equations using instrumental variable procedures. Section 11.4 gives a test of over-identification restrictions whereas, section 11.5 gives a Hausman specification test. Section 11.6 concludes with an empirical example. The Appendix revisits the identification problem and gives a necessary and sufficient condition for identification.

11.1.1 Simultaneous Bias

Example 1: Consider a simple Keynesian model with no government

$$C_t = \alpha + \beta Y_t + u_t \qquad t = 1, 2, \ldots, T \tag{11.1}$$

$$Y_t = C_t + I_t \tag{11.2}$$

where C_t denotes consumption, Y_t denotes disposable income, and I_t denotes autonomous investment. This is a system of two simultaneous equations, also known as *structural equations* with the second equation being an identity. The first equation can be estimated by OLS giving

$$\widehat{\beta}_{OLS} = \textstyle\sum_{t=1}^{T} y_t c_t / \sum_{t=1}^{T} y_t^2 \quad \text{and} \quad \widehat{\alpha}_{OLS} = \bar{C} - \widehat{\beta}_{OLS}\bar{Y} \tag{11.3}$$

with y_t and c_t denoting Y_t and C_t in deviation form, i.e., $y_t = Y_t - \bar{Y}$, and $\bar{Y} = \sum_{t=1}^{T} Y_t/T$. Since I_t is autonomous, it is an *exogenous* variable determined outside the system, whereas C_t and Y_t are *endogenous* variables determined by the system. Let us solve for Y_t and C_t in terms of the constant and I_t. The resulting two equations are known as the *reduced form* equations

$$C_t = \alpha/(1-\beta) + \beta I_t(1-\beta) + u_t/(1-\beta) \tag{11.4}$$

$$Y_t = \alpha/(1-\beta) + I_t/(1-\beta) + u_t/(1-\beta) \tag{11.5}$$

These equations express each endogenous variable in terms of exogenous variables and the error terms. Note that both Y_t and C_t are a function of u_t, and hence both are correlated with u_t. In fact, $Y_t - E(Y_t) = u_t/(1-\beta)$, and

$$\text{cov}(Y_t, u_t) = E[(Y_t - E(Y_t))u_t] = \sigma_u^2/(1-\beta) \geq 0 \quad \text{if} \quad 0 \leq \beta \leq 1 \tag{11.6}$$

This holds because $u_t \sim (0, \sigma_u^2)$ and I_t is exogenous and independent of the error term. Equation (11.6) shows that the right hand side regressor in (11.1) is correlated with the error term. This causes the OLS estimates to be *biased* and *inconsistent*. In fact, from (11.1),

$$c_t = C_t - \bar{C} = \beta y_t + (u_t - \bar{u})$$

and substituting this expression in (11.3), we get

$$\widehat{\beta}_{OLS} = \beta + \textstyle\sum_{t=1}^{T} y_t u_t / \sum_{t=1}^{T} y_t^2 \tag{11.7}$$

From (11.7), it is clear that $E(\widehat{\beta}_{OLS}) \neq \beta$, since the expected value of the second term is not necessarily zero. Also, using (11.5) one gets

$$y_t = Y_t - \bar{Y} = [i_t + (u_t - \bar{u})]/(1-\beta)$$

where $i_t = I_t - \bar{I}$ and $\bar{I} = \sum_{t=1}^{T} I_t/T$. Defining $m_{yy} = \sum_{t=1}^{T} y_t^2/T$, we get

$$m_{yy} = (m_{ii} + 2m_{iu} + m_{uu})/(1-\beta)^2 \tag{11.8}$$

where $m_{ii} = \sum_{t=1}^{T} i_t^2/T$, $m_{iu} = \sum_{t=1}^{T} i_t(u_t - \bar{u})/T$ and $m_{uu} = \sum_{t=1}^{T}(u_t - \bar{u})^2/T$. Also,

$$m_{yu} = (m_{iu} + m_{uu})/(1-\beta) \tag{11.9}$$

Using the fact that plim $m_{iu} = 0$ and plim $m_{uu} = \sigma_u^2$, we get

$$\text{plim } \widehat{\beta}_{OLS} = \beta + \text{plim } (m_{yu}/m_{yy}) = \beta + [\sigma_u^2(1-\beta)/(\text{plim } m_{ii} + \sigma_u^2)]$$

which shows that $\widehat{\beta}_{OLS}$ overstates β if $0 \leq \beta \leq 1$.

Example 2: Consider a simple demand and supply model

$$Q_t^d = \alpha + \beta P_t + u_{1t} \tag{11.10}$$
$$Q_t^s = \gamma + \delta P_t + u_{2t} \tag{11.11}$$
$$Q_t^d = Q_t^s = Q_t \quad t = 1, 2, \ldots, T \tag{11.12}$$

Substituting the equilibrium condition (11.12) in (11.10) and (11.11), we get

$$Q_t = \alpha + \beta P_t + u_{1t} \tag{11.13}$$
$$Q_t = \gamma + \delta P_t + u_{2t} \quad t = 1, 2, \ldots, T \tag{11.14}$$

For the demand equation (11.13), the sign of β is expected to be negative, while for the supply equation (11.14), the sign of δ is expected to be positive. However, we only observe one equilibrium pair (Q_t, P_t) and these are not labeled demand or supply quantities and prices. When we run the OLS regression of Q_t on P_t we do not know what we are estimating, demand or supply? In fact, any linear combination of (11.13) and (11.14) looks exactly like (11.13) or (11.14). It

will have a constant, Price, and a disturbance term in it. Since demand or supply cannot be distinguished from this 'mongrel' we have what is known as an *identification problem*. If the demand equation (or the supply equation) looked different from this mongrel, then this particular equation would be identified. More on this later. For now let us examine the properties of the OLS estimates of the demand equation. It is well known that

$$\widehat{\beta}_{OLS} = \sum_{t=1}^{T} q_t p_t / \sum_{t=1}^{T} p_t^2 = \beta + \sum_{t=1}^{T} p_t(u_{1t} - \bar{u}_1) / \sum_{t=1}^{T} p_t^2 \tag{11.15}$$

where q_t and p_t denote Q_t and P_t in deviation form, i.e., $q_t = Q_t - \bar{Q}$. This estimator is unbiased depending on whether the last term in (11.15) has zero expectations. In order to find this expectation we solve the structural equations in (11.13) and (11.14) for Q_t and P_t

$$Q_t = (\alpha\delta - \gamma\beta)/(\delta - \beta) + (\delta u_{1t} - \beta u_{2t})/(\delta - \beta) \tag{11.16}$$

$$P_t = (\alpha - \gamma)/(\delta - \beta) + (u_{1t} - u_{2t})/(\delta - \beta) \tag{11.17}$$

(11.16) and (11.17) are known as the reduced form equations. Note that both Q_t and P_t are functions of both errors u_1 and u_2. Hence, P_t is correlated with u_{1t}. In fact,

$$p_t = (u_{1t} - \bar{u}_1)/(\delta - \beta) - (u_{2t} - \bar{u}_2)/(\delta - \beta) \tag{11.18}$$

and

$$\text{plim} \sum_{t=1}^{T} p_t(u_{1t} - \bar{u}_1)/T = (\sigma_{11} - \sigma_{12})/(\delta - \beta) \tag{11.19}$$

$$\text{plim} \sum_{t=1}^{T} p_t^2/T = (\sigma_{11} + \sigma_{22} - 2\sigma_{12})/(\delta - \beta)^2 \tag{11.20}$$

where $\sigma_{ij} = \text{cov}(u_{it}, u_{jt})$ for $i, j = 1, 2$; and $t = 1, \ldots, T$. Hence, from (11.15)

$$\text{plim}\ \widehat{\beta}_{OLS} = \beta + (\sigma_{11} - \sigma_{12})(\delta - \beta)/(\sigma_{11} + \sigma_{22} - 2\sigma_{12}) \tag{11.21}$$

and the last term is not necessarily zero, implying that $\widehat{\beta}_{OLS}$ is *not consistent* for β. Similarly, one can show that the OLS estimator for δ is not consistent, see problem 1. This *simultaneous bias* is once again due to the correlation of the right hand side variable (price) with the error term u_1. This correlation could be due to the fact that P_t is a function of u_{2t}, from (11.17), and u_{2t} and u_{1t} are correlated, making P_t correlated with u_{1t}. Alternatively, P_t is a function of Q_t, from (11.13) or (11.14), and Q_t is a function of u_{1t}, from (11.13), making P_t a function of u_{1t}. Intuitively, if a shock in demand (i.e., a change in u_{1t}) shifts the demand curve, the new intersection of demand and supply determines a new equilibrium price and quantity. This new price is therefore, affected by the change in u_{1t}, and is correlated with it.

In general, whenever a right hand side variable is correlated with the error term, the OLS estimates are biased and inconsistent. We refer to this as an *endogeneity problem*. Recall, Figure 3 of Chapter 3 with $\text{cov}(P_t, u_{1t}) > 0$. This shows that P_t's above their mean are on the average associated with u_{1t}'s above their mean, (i.e., $u_{1t} > 0$). This implies that the quantity Q_t associated with this particular P_t is on the average above the true line $(\alpha + \beta P_t)$. This is true for all observations to the right of $E(P_t)$. Similarly, any P_t to the left of $E(P_t)$ is on the average associated with a u_{1t} below its mean, (i.e., $u_{1t} < 0$). This implies that quantities associated with prices below their mean $E(P_t)$ are on the average data points that lie below the true line. With this observed data, the estimated line using OLS will always be biased. In this case, the intercept estimate is biased downwards, whereas the slope estimate is biased upwards. This bias does not

disappear with more data, as any new observation will on the average be either above the true line if $P_t > E(P_t)$ or below the line if $P_t < E(P_t)$. Hence, these OLS estimates are inconsistent.

Deaton (1997, p. 95) has a nice discussion of endogeneity problems in development economics. One important example pertains to farm size and farm productivity. Empirical studies using OLS have found an inverse relationship between productivity as measured by log(Output/Acre) and farm size as measured by (Acreage). This seems counter-intuitive as it suggests that smaller farms are more productive than larger farms. Economic explanations of this phenomenon include the observation that hired labor (which is typically used on large farms) is of lower quality than family labor (which is typically used on small farms). The latter needs less monitoring and can be entrusted with valuable animals and machinery. Another explanation is that this phenomenon is an optimal response by small farmers to uncertainty. It could also be a sign of inefficiency as farmers work too much on their own farms pushing their marginal productivity below market wage. How could this be an endogeneity problem? After all, the amount of land is outside the control of the farmer. This is true, but that does not mean that acreage is uncorrelated with the disturbance term. After all, size is unlikely to be independent of the *quality* of land. "Desert farms that are used for low-intensity animal grazing are typically larger than garden farms, where the land is rich and output/acre is high." In this case, land quality is negatively correlated with land size. It takes more acres to sustain a cow in West Texas than in less arid areas. This negative correlation between acres, the explanatory variable and quality of land which is an omitted variable included in the error term introduces endogeneity. This in turn results in downward bias of the OLS estimate of acreage on productivity.

Endogeneity can also be caused by *sample selection*. Gronau (1973) observed that women with small children had higher wages than women with no children. An economic explanation is that women with children have higher reservation wages and as a result fewer of them work. Of those that work, their observed wages are higher than those without children. The endogeneity works through the unobserved component in the working women's wage that induces her to work. This is positively correlated with the number of children she has and therefore introduces upward biases in the OLS estimate of the effect of the number of children on wages.

11.1.2 The Identification Problem

In general, we can think of any structural equation, say the first, as having one left hand side endogenous variable y_1, g_1 right hand side endogenous variables, and k_1 right hand side exogenous variables. The right hand side endogenous variables are correlated with the error term rendering OLS on this equation biased and inconsistent. Normally, for each endogenous variable, there exists a corresponding structural equation explaining its behavior in the model. We say that a system of simultaneous equations is *complete* if there are as many endogenous variables as there are equations. To correct for the simultaneous bias we need to replace the right hand side endogenous variables in this equation by variables which are highly correlated with the ones they are replacing but not correlated with the error term. Using the method of instrumental variable estimation, discussed below, we will see that these variables turn out to be the predictors obtained by regressing each right hand side endogenous variable on a subset of all the exogenous variables in the system. Let us assume that there are K exogenous variables in the simultaneous system. What set of exogenous variables should we use that would lead to consistent estimates of this structural equation? A search for the minimum set needed for consistency leads us to the *order condition* for identification.

The Order Condition for Identification: A *necessary* condition for identification of any structural equation is that the number of excluded exogenous variables from this equation are greater than or equal to the number of right hand side included endogenous variables. Let K be the number of exogenous variables in the system, then this condition requires $k_2 \geq g_1$, where $k_2 = K - k_1$.

Let us consider the demand and supply equations given in (11.13) and (11.14) but assume that the supply equation has in it an extra variable W_t denoting weather conditions. In this case the demand equation has one right hand side endogenous variable P_t, i.e., $g_1 = 1$ and one excluded exogenous variable W_t, making $k_2 = 1$. Since $k_2 \geq g_1$, this *order condition* is satisfied, in other words, based on the order condition alone we cannot conclude that the demand equation is unidentified. The supply equation, however, has $g_1 = 1$ and $k_2 = 0$, making this equation unidentified, since it does not satisfy the order condition for identification. Note that this condition is only *necessary* but not sufficient for identification. In other words, it is useful only if it is not satisfied, in which case the equation in question is not identified. Note that any linear combination of the new supply and demand equations would have a constant, price and weather. This looks like the supply equation but not like demand. This is why the supply equation is not identified. In order to prove once and for all whether the demand equation is identified, we need the *rank condition* for identification and this will be discussed in details in the Appendix to this chapter. Adding a third variable to the supply equation like the amount of fertilizer used F_t will not help the supply equation any, since a linear combination of supply and demand will still look like supply. However, it does help the identification of the demand equation. Denote by $\ell = k_2 - g_1$, the *degree of over-identification*. In (11.13) and (11.14) both equations are unidentified (or *under-identified*) with $\ell = -1$. When W_t is added to the supply equation, $\ell = 0$ for the demand equation, and it is just-identified. When both W_t and F_t are included in the supply equation, $\ell = 1$ and the demand equation is *over-identified*.

Without the use of matrices, we can describe a two-stage least squares method that will estimate the demand equation consistently. First, we run the right hand side endogenous variable P_t on a constant and W_t and get $\widehat{P}_t$, then replace P_t in the demand equation with $\widehat{P}_t$ and perform this second stage regression. In other words, the first step regression is

$$P_t = \pi_{11} + \pi_{12} W_t + v_t \tag{11.22}$$

with $\widehat{v}_t = P_t - \widehat{P}_t$ satisfying the OLS normal equations $\sum_{t=1}^{T} \widehat{v}_t = \sum_{t=1}^{T} \widehat{v}_t W_t = 0$. The second stage regression is

$$Q_t = \alpha + \beta \widehat{P}_t + \epsilon_t \tag{11.23}$$

with $\sum_{t=1}^{T} \widehat{\epsilon}_t = \sum_{t=1}^{T} \widehat{\epsilon}_t \widehat{P}_t = 0$. Using (11.13) and (11.23), we can write

$$\epsilon_t = \beta(P_t - \widehat{P}_t) + u_{1t} = \beta \widehat{v}_t + u_{1t} \tag{11.24}$$

so that $\sum_{t=1}^{T} \epsilon_t = \sum_{t=1}^{T} u_{1t}$ and $\sum_{t=1}^{T} \epsilon_t \widehat{P}_t = \sum_{t=1}^{T} u_{1t} \widehat{P}_t$ using the fact that $\sum_{t=1}^{T} \widehat{v}_t = \sum_{t=1}^{T} \widehat{v}_t \widehat{P}_t = 0$. So the new error ϵ_t behaves as the original disturbance u_{1t}. However, our right hand side variable is now $\widehat{P}_t$ which is independent of u_{1t} since it is a linear combination of exogenous variables only. We essentially decomposed P_t into two parts, the first part $\widehat{P}_t$ is a linear combination of exogenous variables and therefore, independent of the u_{1t}'s. The second

part is $\widehat{v}_t$ which is correlated with u_{1t}. In fact, this is the source of simultaneous bias. The two parts $\widehat{P}_t$ and $\widehat{v}_t$ are orthogonal to each other by construction. Hence when the $\widehat{v}_t$'s become part of the new error ϵ_t, they are orthogonal to the new regressor $\widehat{P}_t$. Furthermore, $\widehat{P}_t$ is also independent of u_{1t}.

Why would this procedure not work on the estimation of (11.13) if the model is given by equations (11.13) and (11.14). The answer is that in (11.22) we will only have a constant, and no W_t. When we try to run the second-stage regression in (11.23) the regression will fail because of perfect multicollinearity between the constant and $\widehat{P}_t$. This will happen whenever the *order condition* is not satisfied and the equation is not identified, see Kelejian and Oates (1989). Hence, in order for it to succeed in the second stage we need at least one excluded exogenous variable from the demand equation that is in the supply equation, i.e., variables like W_t or F_t. Therefore, whenever the second-stage regression fails because of perfect multicollinearity between the right hand side regressors, this implies that the *order condition* of identification is not satisfied.

In general, if we are given an equation like

$$y_1 = \alpha_{12} y_2 + \beta_{11} X_1 + \beta_{12} X_2 + u_1 \tag{11.25}$$

the order condition requires the existence of at least one exogenous variable excluded from (11.25), say X_3. These extra exogenous variables like X_3 usually appear in other equations of our simultaneous equation model. In the first step regression we run

$$y_2 = \pi_{21} X_1 + \pi_{22} X_2 + \pi_{23} X_3 + v_2 \tag{11.26}$$

with the OLS residuals $\widehat{v}_2$ satisfying

$$\sum_{t=1}^{T} \widehat{v}_{2t} X_{1t} = 0; \quad \sum_{t=1}^{T} \widehat{v}_{2t} X_{2t} = 0; \quad \sum_{t=1}^{T} \widehat{v}_{2t} X_{3t} = 0 \tag{11.27}$$

and in the second step, we run the regression

$$y_1 = \alpha_{12} \widehat{y}_2 + \beta_{11} X_1 + \beta_{12} X_2 + \epsilon_1 \tag{11.28}$$

where $\epsilon_1 = \alpha_{12}(y_2 - \widehat{y}_2) + u_1 = \alpha_{12}\widehat{v}_2 + u_1$. This regression will lead to consistent estimates, because

$$\begin{aligned} &\sum_{t=1}^{T} \widehat{y}_{2t} \epsilon_{1t} = \sum_{t=1}^{T} \widehat{y}_{2t} u_{1t}; \quad \sum_{t=1}^{T} X_{1t} \epsilon_{1t} = \sum_{t=1}^{T} X_{1t} u_{1t}; \\ &\sum_{t=1}^{T} X_{2t} \epsilon_{1t} = \sum_{t=1}^{T} X_{2t} u_{1t} \end{aligned} \tag{11.29}$$

and u_{1t} is independent of the exogenous variables. In order to solve for 3 structural parameters α_{12}, β_{11} and β_{12} one needs three linearly independent OLS normal equations. $\sum_{t=1}^{T} \widehat{y}_{2t} \widehat{\epsilon}_{1t} = 0$ is a new piece of information provided y_2 is regressed on at least one extra variable besides X_1 and X_2. Otherwise, $\sum_{t=1}^{T} X_{1t} \widehat{\epsilon}_{1t} = \sum_{t=1}^{T} X_{2t} \widehat{\epsilon}_{1t} = 0$ are the only two linearly independent normal equations in three structural parameters.

What happens if there is another right hand side endogenous variable, say y_3? In that case (11.25) becomes

$$y_1 = \alpha_{12} y_2 + \alpha_{13} y_3 + \beta_{11} X_1 + \beta_{12} X_2 + u_1 \tag{11.30}$$

Now we need at least two exogenous variables that are excluded from (11.30) for the order condition to be satisfied, and the second stage regression to run. Otherwise, we will have less

linearly independent equations than there are structural parameters to estimate, and the second stage regression will fail. Also, y_2 and y_3 should be regressed on the *same set* of exogenous variables. Furthermore, this set of second-stage regressors *should always include the right hand side* exogenous variables of (11.30). These two conditions will ensure consistency of the estimates. Let X_3 and X_4 be the excluded exogenous variables from (11.30). Our first step regression would regress y_2 and y_3 on X_1, X_2, X_3 and X_4 to get $\widehat{y}_2$ and $\widehat{y}_3$, respectively. The second stage regression would regress y_1 on $\widehat{y}_2$, $\widehat{y}_3$, X_1 and X_2. From the first step regressions we have

$$y_2 = \widehat{y}_2 + \widehat{v}_2 \quad \text{and} \quad y_3 = \widehat{y}_3 + \widehat{v}_3 \tag{11.31}$$

where $\widehat{y}_2$ and $\widehat{y}_3$ are linear combinations of the X's, and $\widehat{v}_2$ and $\widehat{v}_3$ are the residuals. The second stage regression has the following normal equations

$$\sum_{t=1}^{T} \widehat{y}_{2t}\widehat{\epsilon}_{1t} = \sum_{t=1}^{T} \widehat{y}_{3t}\widehat{\epsilon}_{1t} = \sum_{t=1}^{T} X_{1t}\widehat{\epsilon}_{1t} = \sum_{t=1}^{T} X_{2t}\widehat{\epsilon}_{1t} = 0 \tag{11.32}$$

where $\widehat{\epsilon}_1$ denotes the residuals from the second stage regression. In fact

$$\epsilon_1 = \alpha_{12}\widehat{v}_2 + \alpha_{13}\widehat{v}_3 + u_1 \tag{11.33}$$

Now $\sum_{t=1}^{T} \epsilon_{1t}\widehat{y}_{2t} = \sum_{t=1}^{T} u_{1t}\widehat{y}_{2t}$ because $\sum_{t=1}^{T} \widehat{v}_{2t}\widehat{y}_{2t} = \sum_{t=1}^{T} \widehat{v}_{3t}\widehat{y}_{2t} = 0$. The latter holds because $\widehat{y}_2$, the predictor, is orthogonal to $\widehat{v}_2$, the residual. Also, $\widehat{y}_2$ is orthogonal to $\widehat{v}_3$ if y_2 is regressed on a set of X's that are a subset of the regressors included in the first step regression of y_3. Similarly, $\sum_{t=1}^{T} \epsilon_{1t}\widehat{y}_{3t} = \sum_{t=1}^{T} u_{1t}\widehat{y}_{3t}$ if y_3 is regressed on a set of exogenous variables that are a subset of the X's included in the first step regression of y_2. Combining these two conditions leads to the following fact: y_2 and y_3 have to be regressed on the *same set* of exogenous variables for the composite error term to behave like the original error. Furthermore these exogenous variables *should include the included X's* on the right hand side of the equation to be estimated, i.e., X_1 and X_2, otherwise, $\sum_{t=1}^{T} \epsilon_{1t}X_{1t}$ is not necessarily equal to $\sum_{t=1}^{T} u_{1t}X_{1t}$, because $\sum_{t=1}^{T} \widehat{v}_{2t}X_{1t}$ or $\sum_{t=1}^{T} \widehat{v}_{3t}X_{1t}$ are not necessarily zero. For further analysis along these lines, see problem 2.

11.2 Single Equation Estimation: Two-Stage Least Squares

In matrix form, we can write the first structural equation as

$$y_1 = Y_1\alpha_1 + X_1\beta_1 + u_1 = Z_1\delta_1 + u_1 \tag{11.34}$$

where y_1 and u_1 are $(T \times 1)$, Y_1 denotes the right hand side endogenous variables which is $(T \times g_1)$ and X_1 is the set of right hand side included exogenous variables which is $(T \times k_1)$, α_1 is of dimension g_1 and β_1 is of dimension k_1. $Z_1 = [Y_1, X_1]$ and $\delta_1' = (\alpha_1', \beta_1')$. We require the existence of excluded exogenous variables, from (11.34), call them X_2, enough to identify this equation. These excluded exogenous variables appear in the other equations in the simultaneous model. Let the set of all exogenous variables be $X = [X_1, X_2]$ where X is of dimension $(T \times k)$. For the order condition to be satisfied for equation (11.34) we must have $(k - k_1) \geq g_1$. If *all* the exogenous variables in the system are included in the first step regression, i.e., Y_1 is regressed on X to get $\widehat{Y}_1$, the resulting second stage least squares estimator obtained from regressing y_1 on $\widehat{Y}_1$ and X_1 is called two-stage least squares (2SLS). This method was proposed independently by Basmann (1957) and Theil (1953). In matrix form $\widehat{Y}_1 = P_X Y_1$ is the predictor of the right

hand side endogenous variables, where P_X is the projection matrix $X(X'X)^{-1}X'$. Replacing Y_1 by $\widehat{Y}_1$ in (11.34), we get

$$y_1 = \widehat{Y}_1\alpha_1 + X_1\beta_1 + w_1 = \widehat{Z}_1\delta_1 + w_1 \tag{11.35}$$

where $\widehat{Z}_1 = [\widehat{Y}_1, X_1]$ and $w_1 = u_1 + (Y_1 - \widehat{Y}_1)\alpha_1$. Running OLS on (11.35) one gets

$$\widehat{\delta}_{1,2SLS} = (\widehat{Z}_1'\widehat{Z}_1)^{-1}\widehat{Z}_1'y_1 = (Z_1'P_XZ_1)^{-1}Z_1'P_Xy_1 \tag{11.36}$$

where the second equality follows from the fact that $\widehat{Z}_1 = P_XZ_1$ and the fact that P_X is idempotent. The former equality holds because $P_XX = X$, hence $P_XX_1 = X_1$, and $P_XY_1 = \widehat{Y}_1$. If there is only one right hand side endogenous variable, running the first-stage regression y_2 on X_1 and X_2 and testing that the coefficients of X_2 are *all* zero against the hypothesis that at least one of these coefficients is different from zero is a test for rank identification. In case of several right hand side endogenous variables, things get complicated, see Cragg and Donald (1996), but one can still run the first-stage regressions for each right hand side endogenous variable to make sure that at least one element of X_2 is significantly different from zero.[1] This is not sufficient for the rank condition but it is a good diagnostic for whether the rank condition fails. If we fail to meet this requirement we should question our 2SLS estimator.

Two-stage least squares can also be thought of as a simple instrumental variables estimator with the set of instruments $W = \widehat{Z}_1 = [\widehat{Y}_1, X_1]$. Recall that Y_1 is correlated with u_1, rendering OLS inconsistent. The idea of simple instrumental variables is to find a set of instruments, say W for Z_1 with the following properties: (1) plim $W'u_1/T = 0$, the instruments have to be *exogenous*, i.e., uncorrelated with the error term, otherwise this defeats the purpose of the instruments and result in inconsistent estimates. (2) plim $W'W/T = Q_w \neq 0$, where Q_w is finite and positive definite, the W's should not be perfectly multicollinear. (3) W should be highly correlated with Z_1, i.e., the instruments should be *highly relevant*, not *weak instruments* as we will explain shortly. In fact, plim $W'Z_1/T$ should be finite and of full rank $(k_1 + g_1)$. Premultiplying (11.34) by W', we get

$$W'y_1 = W'Z_1\delta_1 + W'u_1 \tag{11.37}$$

In this case, $W = \widehat{Z}_1$ is of the same dimension as Z_1, and since plim $W'Z_1/T$ is square and of full rank $(k_1 + g_1)$, the *simple instrumental variable* (IV) estimator of δ_1 becomes

$$\widehat{\delta}_{1,IV} = (W'Z_1)^{-1}W'y_1 = \delta_1 + (W'Z_1)^{-1}W'u_1 \tag{11.38}$$

with plim $\widehat{\delta}_{1,IV} = \delta_1$ which follows from (11.37) and the fact that plim $W'u_1/T = 0$.

Digression: In the general linear model, $y = X\beta + u$, X is the set of instruments for X. Premultiplying by X' we get $X'y = X'X\beta + X'u$ and using the fact that plim $X'u/T = 0$, one gets

$$\widehat{\beta}_{IV} = (X'X)^{-1}X'y = \widehat{\beta}_{OLS}.$$

This estimator is consistent as long as X and u are uncorrelated. In the simultaneous equation model for the first structural equation given in (11.34), the right hand side regressors Z_1 include endogenous variables Y_1 that are correlated with u_1. Therefore OLS on (11.34) will lead to

inconsistent estimates, since the matrix of instruments $W = Z_1$, and Z_1 is correlated with u_1. In fact,

$$\widehat{\delta}_{1,OLS} = (Z_1'Z_1)^{-1}Z_1'y_1 = \delta_1 + (Z_1'Z_1)^{-1}Z_1'u_1$$

with plim $\widehat{\delta}_{1,OLS} \neq \delta_1$ since plim $Z_1'u_1/T \neq 0$.

Denote by $e_{1,OLS} = y_1 - Z_1\widehat{\delta}_{1,OLS}$ as the OLS residuals on the first structural equation, then

$$\text{plim } s_1^2 = \frac{e_{1,OLS}'e_{1,OLS}}{T-(g_1+k_1)} = \text{plim}\frac{u_1'\bar{P}_{Z_1}u_1}{T-(g_1+k_1)} = \sigma_{11} - \text{plim}\frac{u_1'Z_1(Z_1'Z_1)^{-1}Z_1'u_1}{T-(g_1+k_1)} \leq \sigma_{11},$$

since the last term is positive. Only if plim $Z_1'u_1/T$ is zero will plim $s_1^2 = \sigma_{11}$, otherwise it is smaller. OLS fits very well, it minimizes $(y_1 - Z_1\delta_1)'(y_1 - Z_1\delta_1)$. Since Z_1 and u_1 are correlated, OLS attributes part of the variation in y_1 that is due to u_1 incorrectly to the regressor Z_1.

Both the simple IV and OLS estimators can be interpreted as method of moments estimators. These were discussed in Chapter 2. For OLS, the population moment conditions are given by $E(X'u) = 0$ and the corresponding sample moment conditions yield $X'(y - X\widehat{\beta})/T = 0$. Solving for $\widehat{\beta}$ results in $\widehat{\beta}_{OLS}$. Similarly, the population moment conditions for the simple IV estimator in (11.37) are $E(W'u_1) = 0$ and the corresponding sample moment conditions yield $W'(y_1 - Z_1\widehat{\delta}_1)/T = 0$. Solving for $\widehat{\delta}_1$ results in $\widehat{\delta}_{1,IV}$ given in (11.38).

If $W = [\widehat{Y}_1, X_1]$, then (11.38) results in

$$\widehat{\delta}_{1,IV} = \begin{bmatrix} \widehat{Y}_1'Y_1 & \widehat{Y}_1'X_1 \\ X_1'Y_1 & X_1'X_1 \end{bmatrix}^{-1} \begin{bmatrix} \widehat{Y}_1'y_1 \\ X_1'y_1 \end{bmatrix} \tag{11.39}$$

which is the same as (11.36)

$$\widehat{\delta}_{1,2SLS} = \begin{bmatrix} \widehat{Y}_1'\widehat{Y}_1 & \widehat{Y}_1'X_1 \\ X_1'\widehat{Y}_1 & X_1'X_1 \end{bmatrix}^{-1} \begin{bmatrix} \widehat{Y}_1'y_1 \\ X_1'y_1 \end{bmatrix} \tag{11.40}$$

provided $\widehat{Y}_1'\widehat{Y}_1 = \widehat{Y}_1'Y_1$, and $X_1'Y_1 = X_1'\widehat{Y}_1$. The latter conditions hold because $\widehat{Y}_1 = P_XY_1$, and $P_XX_1 = X_1$.

In general, let X^* be our set of first stage regressors. An IV estimator with $\widehat{Y}_1^* = P_{X^*}Y_1$, i.e., with every right hand side y regressed on the *same set of regressors* X^*, will satisfy

$$\widehat{Y}_1^{*\prime}\widehat{Y}_1^* = \widehat{Y}_1^{*\prime}P_{X^*}Y_1 = \widehat{Y}_1^{*\prime}Y_1$$

In addition, for $X_1'\widehat{Y}_1^*$ to equal $X_1'Y_1$, X_1 has to be a subset of the regressors in X^*. Therefore X^* *should include* X_1 *and at least as many* X*'s from* X_2 *as is required for identification*, i.e., (at least g_1 of the X's from X_2). In this case, the IV estimator using $W^* = [\widehat{Y}_1^*, X_1]$ will result in the same estimator as that obtained by a two stage regression where in the first step $\widehat{Y}_1^*$ is obtained by regressing Y_1 on X^*, and in the second step y_1 is regressed on W^*. Note that these are the same conditions required for consistency of an IV estimator. Note also, that if this equation is just-identified, then there is exactly g_1 of the X's excluded from that equation. In other words, X_2 is of dimension $(T \times g_1)$, and $X^* = X$ is of dimension $T \times (g_1 + k_1)$. Problem 3 shows that 2SLS in this case reduces to an IV estimator with $W = X$, i.e.

$$\widehat{\delta}_{1,2SLS} = \widehat{\delta}_{1,IV} = (X'Z_1)^{-1}X'y_1 \tag{11.41}$$

Note that if the first equation is over-identified, then $X'Z_1$ is not square and (11.41) cannot be computed.

Rather than having W, the matrix of instruments, be of exactly the same dimension as Z_1 which is required for the expression in (11.38), one can define *a generalized instrumental variable* in terms of a general matrix W of dimension $T \times \ell$ where $\ell \geq g_1 + k_1$. The latter condition is the *order condition* for identification. In this case, $\widehat{\delta}_{1,IV}$ is obtained as GLS on (11.37). Using the fact that

$$\text{plim } W'u_1u_1'W/T = \sigma_{11} \text{ plim } W'W/T,$$

one gets

$$\widehat{\delta}_{1,IV} = (Z_1'P_WZ_1)^{-1}Z_1'P_Wy_1 = \delta_1 + (Z_1'P_WZ_1)^{-1}Z_1'P_Wu_1$$

with plim $\widehat{\delta}_{1,IV} = \delta_1$ and limiting covariance matrix σ_{11} plim $(Z_1'P_WZ_1/T)^{-1}$. Therefore, 2SLS can be obtained as a generalized instrumental variable estimator with $W = X$. This also means that 2SLS of δ_1 can be obtained as GLS on (11.34) after premultiplication by X', see problem 4. Note that GLS on (11.37) minimizes $(y_1 - Z_1\delta_1)'P_W(y_1 - Z_1\delta_1)$ which yields the first-order conditions

$$Z_1'P_W(y_1 - Z_1\widehat{\delta}_{1,IV}) = 0$$

the solution of which is $\widehat{\delta}_{1,IV} = (Z_1'P_WZ_1)^{-1}Z_1'P_Wy_1$. It can also be shown that 2SLS and the generalized instrumental variables estimators are special cases of a Generalized Method of Moments (GMM) estimator considered by Hansen (1982). See Davidson and MacKinnon (1993) and Hall (1993) for an introduction to GMM.

For the matrix $Z_1'P_WZ_1$ to be of full rank and invertible, a *necessary* condition is that W must be of full rank $\ell \geq (g_1 + k_1)$. This is in fact, the *order condition* of identification. If $\ell = g_1 + k_1$, then this equation is just-identified. Also, $W'Z_1$ is square and nonsingular. Problem 10 asks the reader to verify that the generalized instrumental variable estimator reduces to the simple instrumental variable estimator given in (11.38). Also, under just-identification the minimized value of the criterion function is zero.

One of the biggest problems with IV estimation is the choice of the instrumental variables W. We have listed some necessary conditions for this set of instruments to yield consistent estimators of the structural coefficients. However, different choices by different researchers may yield different estimates in finite samples. Using more instruments will yield more efficient IV estimation. Let W_1 and W_2 be two sets of IV's with W_1 being spanned by the space of W_2. In this case, $P_{W_2}W_1 = W_1$ and therefore, $P_{W_2}P_{W_1} = P_{W_1}$. For the corresponding IV estimators

$$\widehat{\delta}_{1,W_i} = (Z_1'P_{W_i}Z_1)^{-1}Z_1'P_{W_i}y_1 \quad \text{for } i = 1, 2$$

are both consistent for δ_1 as long as plim $W_i'u_1/T = 0$ and have asymptotic covariance matrices

$$\sigma_{11} \text{ plim } (Z_1'P_{W_i}Z_1/T)^{-1}$$

Note that $\widehat{\delta}_{1,W_2}$ is at least as efficient as $\widehat{\delta}_{1,W_1}$, if the difference in their asymptotic covariance matrices is positive semi-definite, i.e., if

$$\sigma_{11}\left[\text{plim}\frac{Z_1'P_{W_1}Z_1}{T}\right]^{-1} - \sigma_{11}\left[\text{plim}\frac{Z_1'P_{W_2}Z_1}{T}\right]^{-1}$$

is p.s.d. This holds, if $Z_1'P_{W_2}Z_1 - Z_1'P_{W_1}Z_1$ is p.s.d. This last condition holds since $P_{W_2} - P_{W_1}$ is idempotent. Problem 11 asks the reader to verify this result. $\widehat{\delta}_{1,W_2}$ is more efficient than $\widehat{\delta}_{1,W_1}$ since W_2 explains Z_1 at least as well as W_1. This seems to suggest that one should use as many instruments as possible. If T is large this is a good strategy. But, if T is finite, there will be a trade-off between this gain in asymptotic efficiency and the introduction of more finite sample bias in our IV estimator.

In fact, the more instruments we use, the more will $\widehat{Y}_1$ resemble Y_1 and the more bias is introduced in this second stage regression. The extreme case where Y_1 is perfectly predicted by $\widehat{Y}_1$ returns us to OLS which we know is biased. On the other hand, if our set of instruments have little ability in predicting Y_1, then the resulting instrumental variable estimator will be inefficient and its asymptotic distribution will not resemble its finite sample distribution, see Nelson and Startz (1990). If the number of instruments is fixed and the coefficients of the instruments in the first stage regression go to zero at the rate $1\sqrt{T}$, indicating weak correlation, Staiger and Stock (1997) find that even as T increases, IV estimation is not consistent and has a nonstandard asymptotic distribution. Bound et al. (1995) recommend reporting the R^2 or the F-statistic of the first stage regression as a useful indicator of the quality of IV estimates.

Instrumental variables are important for obtaining consistent estimates when endogeneity is suspected. However, invalid instruments can produce meaningless results. How do we know whether our instruments are valid? Stock and Watson (2003) draw an analogy between a relevant instrument and a large sample. The more relevant the instrument, i.e., the more the variation in the right hand side endogenous variable that is explained by this instrument, the more accurate the resulting estimator. This is similar to the observation that the larger the sample size, the more accurate the estimator. They argue that the instruments should not just be relevant, but highly relevant if the normal distribution is to provide a good approximation to the sampling distribution of 2SLS. Weak instruments explain little of the variation in the right hand side endogenous variable they are instrumenting. This renders the normal distribution as a poor approximation to the sampling distribution of 2SLS, even if the sample size is large. Stock and Watson (2003) suggest a simple rule of thumb to check for weak instruments. If there is one right hand side endogenous variable, the first-stage regression can test for the significance of the excluded exogenous variables (or instruments) using an F-statistic. This first-stage F-statistic should be larger than 10.[2] Stock and Watson (2003) suggest that a first-stage F-statistic less than 10 indicates weak instruments which casts doubt on the validity of 2SLS, since with weak instruments, 2SLS will be biased even in large samples and the corresponding t-statistics and confidence intervals will be unreliable. Finding weak instruments, one can search for additional stronger instruments, or use alternative estimators than 2SLS which are less sensitive to weak instruments like LIML. Deaton (1997, p. 112) argues that it is difficult to find instruments that are exogenous while at the same time highly correlated with the endogenous variables they are instrumenting. He argues that it is easy to generate 2SLS estimates that are different from OLS but much harder to make the case that these 2SLS estimates are necessarily better than OLS. "Credible identification and estimation of structural equations almost always requires real creativity, and creativity cannot be reduced to a formula." Stock and Watson (2003, p. 371) show that for the case of a single right hand side endogenous variable with no included exogenous variables and one weak instrument, the distribution of the 2SLS estimators is non-normal even for large samples, with the mean of the sampling distribution of the 2SLS estimator approximately equal to the true coefficient plus the asymptotic bias of the OLS estimator divided

by $(E(F)-1)$ where F is the first-stage F-statistic. If $E(F)=10$, then the large sample bias of 2SLS is (1/9) that of the large sample bias of OLS. They argue that this rule of thumb is an acceptable cutoff for most empirical applications.

2SLS is a single equation estimator. The focus is on a particular equation. $[y_1, Y_1, X_1]$ is specified and therefore all that is needed to perform 2SLS is the matrix X of all exogenous variables in the system. If a researcher is interested in a particular behavioral economic relationship which may be a part of a big model consisting of several equations, one need not specify the whole model to perform 2SLS on that equation, all that is needed is the matrix of all exogenous variables in that system. Empirical studies involving one structural equation, specify which right hand side variables are endogenous and proceed by estimating this equation via an IV procedure that usually includes all the feasible exogenous variables available to the researcher. If this set of exogenous variables does not include all the X's in the system, this estimation method is not 2SLS. However, it is a consistent IV method which we will call feasible 2SLS.

Substituting (11.34) in (11.36), we get

$$\widehat{\delta}_{1,2SLS} = \delta_1 + (Z_1' P_X Z_1)^{-1} Z_1' P_X u_1 \tag{11.42}$$

with plim $\widehat{\delta}_{1,2SLS} = \delta_1$ and an asymptotic variance covariance matrix given by σ_{11} plim $(Z_1' P_X Z_1/T)^{-1}$. σ_{11} is estimated from the 2SLS residuals $\widehat{u}_1 = y_1 - Z_1 \widehat{\delta}_{1,2SLS}$, by computing $s_{11} = \widehat{u}_1' \widehat{u}_1/(T - g_1 - k_1)$. It is important to emphasize that s_{11} is obtained from the 2SLS residuals of the original equation (11.34), not (11.35). In other words, s_{11} is not the mean squared error (i.e., s^2) of the second stage regression given in (11.35). The latter regression has $\widehat{Y}_1$ in it and not Y_1. Therefore, the asymptotic variance covariance matrix of 2SLS can be estimated by $s_{11}(Z_1' P_X Z_1)^{-1} = s_{11}(\widehat{Z}_1' \widehat{Z}_1)^{-1}$. The t-statistics reported by 2SLS packages are based on the standard errors obtained from the square root of the diagonal elements of this matrix. These standard errors and t-statistics can be made robust for heteroskedasticity by computing $(\widehat{Z}_1' \widehat{Z}_1)^{-1}(\widehat{Z}_1' diag[\widehat{u}_i^2] \widehat{Z}_1)(\widehat{Z}_1' \widehat{Z}_1)^{-1}$ where $\widehat{u}_i$ denotes the i-th 2SLS residual. Wald type statistics for $H_o; R\delta_1 = r$ based on 2SLS estimates of δ_1 can be obtained as in equation (7.41) with $\widehat{\delta}_{1,2SLS}$ replacing $\widehat{\beta}_{OLS}$ and $var(\widehat{\delta}_{1,2SLS}) = s_{11}(\widehat{Z}_1' \widehat{Z}_1)^{-1}$ replacing $var(\widehat{\beta}_{OLS}) = s_{11}(X'X)^{-1}$. This can be made robust for heteroskedasticity by using the robust variance covariance matrix of $\widehat{\delta}_{1,2SLS}$ described above. The resulting Wald statistic is asymptotically distributed as χ^2_q under the null hypothesis, with q being the number of restrictions imposed by $R\delta_1 = r$.

LM type tests for exclusion restrictions, like a subset of δ_1 set equal to zero can be performed by running the *restricted* 2SLS residuals on the matrix of *unrestricted* second stage regressors $\widehat{Z}_1$. The test statistic is given by TR_u^2 where R_u^2 denotes the uncentered R^2. This is asymptotically distributed as χ^2_q under the null hypothesis, where q is the number of coefficients in δ_1 set equal to zero. Note that it does not matter whether the exclusion restrictions are imposed on β_1 or α_1, i.e., whether the excluded variables to be tested are endogenous or exogenous. An F-test for these exclusion restrictions can be constructed based on the restricted and unrestricted residual sums of squares from the second stage regression. The denominator of this F-statistic, however, is based on the unrestricted 2SLS residual sum of squares as reported by the 2SLS package. Of course, one has to adjust the numerator and denominator by the appropriate degrees of freedom. Under the null, this is asymptotically distributed as $F(q, T-(g_1+k_1))$. See Wooldridge (1990) for details. Also, see the over-identification test in Section 11.5.

Finite sample properties of 2SLS are model specific, see Mariano (2001) for a useful summary. One important result is that the absolute moments of positive order for 2SLS are finite up to the order of over-identification. So, for the 2SLS estimator to have a mean and variance, we

need the degree of over-identification to be at least 2. This also means that for a just-identified model, no moments for 2SLS exist. For 2SLS, the absolute bias is an increasing function of the degree of over-identification. For the case of one right hand side included endogenous regressor, like equation (11.25), the size of OLS bias relative to 2SLS gets larger, the lower the degree of over-identification, the bigger the sample size, the higher the absolute value of the correlation between the disturbances and the endogenous regressor y_2 and the higher the *concentration parameter* μ^2. The latter is defined as $\mu^2 = E(y_2)'(P_X - P_{X_1})E(y_2)/\omega^2$ and $\omega^2 = \text{var}(y_{2t})$. In terms of MSE, larger values of μ^2 and large sample size favor 2SLS over OLS.

Another important single equation estimator is the *Limited Information Maximum Likelihood* (LIML) estimator which as the name suggests maximizes the likelihood function pertaining to the endogenous variables appearing in the estimated equation only. Excluded exogenous variables from this equation as well as the identifiability restrictions on other equations in the system are disregarded in the likelihood maximization. For details, see Anderson and Rubin (1950). LIML is invariant to the normalization choice of the dependent variable whereas 2SLS is not. This invariancy of LIML is in the spirit of a simultaneous equation model where normalization should not matter. Under just-identification 2SLS and LIML are equivalent. LIML is also known as the *Least Variance Ratio* (LVR) method, since the LIML estimates can be obtained by minimizing a ratio of two variances or equivalently the ratio of two residual sum of squares. Using equation (11.34), one can write

$$y_1^* = y_1 - Y_1\alpha = X_1\beta_1 + u_1$$

For a choice of α one can compute y_1^* and regress it on X_1 to get the residual sum of squares RSS_1. Now regress y_1^* on X_1 and X_2 and compute the residual sum of squares RSS_2. Equation (11.34) states that X_2 does not enter the specification of that equation. In fact, this is where our identifying restrictions come from and the excluded exogenous variables that are used as instrumental variables. If these identifying restrictions are true, adding X_2 to the regression of y_1^* and X_1 should lead to minimal reduction in RSS_1. Therefore, the LVR method finds the α that will minimize the ratio (RSS_1/RSS_2). After α is estimated, β_1 is obtained from regressing y_1^* on X_1. In contrast, it can be shown that 2SLS minimizes $RSS_1 - RSS_2$. For details, see Johnston (1984) or Mariano (2001). Estimator bias is less of a problem for LIML than 2SLS. In fact as the number of instruments increase with the sample size such that their ratio is a constant, Bekker (1994) shows that 2SLS becomes inconsistent while LIML remains consistent.

Example 3: Simple Keynesian Model

For the data from the Economic Report of the President, given in Table 5.1, consider the simple Keynesian model with no government

$$C_t = \alpha + \beta Y_t + u_t \quad t = 1, 2, \ldots, T$$

with $Y_t = C_t + I_t$.

The OLS estimates of the consumption function yield:

$$\begin{array}{lll} C_t = -65.79 + & 0.916 & Y_t \\ \quad (90.99) & (0.009) & \end{array}$$

The 2SLS estimates assuming that I_t is exogenous and is the only instrument available, yield

$$\begin{array}{lll} C_t = 313.01 + & 0.878 & Y_t \\ \quad (129.8) & (0.012) & \end{array}$$

Table 11.1 Two-Stage Least Squares

Dependent Variable:	CONSUM			
Method:	Two-Stage Least Squares			
Sample:	1950 1993			
Included observations:	44			
Instrument list:	I			
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	313.0072	129.8318	2.410867	0.0204
Y	0.878394	0.012464	70.47380	0.0000
R-squared	0.994620	Mean dependent var		9250.545
Adjusted R-squared	0.994492	S.D. dependent var		2484.624
S.E. of regression	184.4002	Sum squared resid		1428145
F-statistic	4966.557	Durbin-Watson stat		0.31203
Prob (F-statistic)	0.00000			

Table 11.1 reports these 2SLS results using EViews. Note that the OLS estimate of the intercept is understated, while that of the slope estimate is overstated indicating positive correlation between Y_t and the error as described in (11.6).

OLS on the reduced form equations yield

$$\underset{(811.9)}{C_t = 2573.94 +} \underset{(0.843)}{7.22} \; I_t \quad \text{and} \quad \underset{(811.9)}{Y_t = 2573.94 +} \underset{(0.843)}{8.22} \; I_t$$

From example (A.5) in the Appendix, we see that $\widehat{\beta} = \widehat{\pi}_{12}/\widehat{\pi}_{22} = 7.22/8.22 = 0.878$ as described in (A.24). Also, $\widehat{\beta} = (\widehat{\pi}_{22} - 1)/\widehat{\pi}_{22} = (8.22 - 1)/8.22 = 7.22/822 = 0.878$ as described in (A.25). Similarly, $\widehat{\alpha} = \widehat{\pi}_{11}/\widehat{\pi}_{22} = \widehat{\pi}_{21}/\widehat{\pi}_{22} = 2573.94/8.22 = 313.01$ as described in (A.22).

This confirms that under just-identification, the 2SLS estimates of the structural coefficients are identical to the *Indirect Least Squares* (ILS) estimates. The latter estimates uniquely solve for the structural parameter estimates from the reduced form estimates under just-identification. Note that in this case both 2SLS and ILS estimates of the consumption equation are identical to the simple IV estimator using I_t as an instrument for Y_t; i.e., $\widehat{\beta}_{IV} = m_{ci}/m_{yi}$ as shown in (A.24).

11.2.1 Spatial Lag Dependence

An alternative popular model for spatial lag dependence considered in Section 9.9 is given by:

$$y = \rho W y + X\beta + \epsilon$$

where $\epsilon \sim \text{IIN}(0, \sigma^2)$, see Anselin (1988). Here y_i may denote output in region i which is affected by output of its neighbors through the spatial coefficient ρ and the weight matrix W. Recall from section 9.9, W is a *known* weight matrix with zero elements along its diagonal. It could be a contiguity matrix having elements 1 if its a neighboring region and zero otherwise. Usually this is normalized such that each row sums to 1. Alternatively, W could be based on distances

from neighbors again normalized such that each row sums to 1. It is clear that the presence of Wy as a regressor introduces *endogeneity*. Assuming $(I_n - \rho W)$ nonsingular, one can solve for the reduced form model:

$$y = (I_n - \rho W)^{-1} X\beta + \epsilon^*$$

where $\epsilon^* = (I_n - \rho W)^{-1}\epsilon$ has mean zero and variance covariance matrix which has the same form as (9.38), i.e.,

$$\Sigma = E(\epsilon^* \epsilon^{*\prime}) = \sigma^2 \Omega = \sigma^2 (I_n - \rho W)^{-1} (I_n - \rho W')^{-1}$$

For $|\rho| < 1$, one obtains

$$(I_n - \rho W)^{-1} = I_n + \rho W + \rho^2 W^2 + \rho^3 W^3 + ...$$

Hence

$$E(y/X) = (I_n - \rho W)^{-1} X\beta = X\beta + \rho W X\beta + \rho^2 W^2 X\beta + \rho^3 W^3 X\beta + ...$$

This also means that

$$E(Wy/X) = W(I_n - \rho W)^{-1} X\beta = WX\beta + \rho W^2 X\beta + \rho^2 W^3 X\beta + \rho^3 W^4 X\beta + ...$$

Based on this last expression, Kelejian and Robinson (1993) and Kelejian and Prucha (1998) suggest the use of a subset of the following instrumental variables:

$$\{X, WX, W^2X, W^3X, W^4X, ...\}$$

Lee (2003) suggested using the optimal instrument matrix:

$$\{X, W(I_n - \widehat{\rho} W)^{-1} X\widehat{\beta}\}$$

where the values for $\widehat{\rho}$ and $\widehat{\beta}$ are obtained from a first stage IV estimator, using $\{X, WX\}$ as instruments, possibly augmented with W^2X. Note that Lee's (2003) instruments involve inverting a matrix of dimension n. Kelejian, et al. (2004) suggest an approximation based upon:

$$\{X, \sum_{s=0}^{r} \widehat{\rho}^s W^{s+1} X\widehat{\beta}\}$$

where r, the highest order of this approximation depends upon the sample size, with $r = o(n^{1/2})$. In their Monte Carlo experiments, they set $r = n^c$ where $c = 0.25, 0.35$, and 0.45. This is a natural application of 2SLS to deal with the problem of spatial lag dependence.

11.3 System Estimation: Three-Stage Least Squares

If the entire simultaneous equations model is to be estimated, then one should consider system estimators rather than single equation estimators. System estimators take into account the zero restrictions in every equation as well as the variance-covariance matrix of the disturbances of

the whole system. One such system estimator is *Three-Stage Least Squares* (3SLS) where the structural equations are stacked on top of each other, just like a set of SUR equations,

$$y = Z\delta + u \tag{11.43}$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_G \end{bmatrix};\ Z = \begin{bmatrix} Z_1 & 0 & \dots & 0 \\ 0 & Z_2 & \dots & 0 \\ \vdots & \dots & \vdots & \dots \\ 0 & \dots & & Z_G \end{bmatrix};\ \delta = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_G \end{bmatrix};\ u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_G \end{bmatrix}$$

and u has zero mean and variance-covariance matrix $\Sigma \otimes I_T$, indicating the possible correlation among the disturbances of the different structural equations. $\Sigma = [\sigma_{ij}]$, with $E(u_i u_j') = \sigma_{ij} I_T$, for $i, j = 1, 2, \dots, G$. This $\otimes$ notation was used in Chapter 9 and defined in the Appendix to Chapter 7. Problem 4 shows that premultiplying the i-th structural equation by X' and performing GLS on the transformed equation results in 2SLS. For the system given in (11.43), the analogy is obtained by premultiplying by $(I_G \otimes X')$, i.e., each equation by X', and performing GLS on the whole system. The transformed error $(I_G \otimes X')u$ has a zero mean and variance-covariance matrix $\Sigma \otimes (X'X)$. Hence, GLS on the entire system obtains

$$\begin{aligned} \widehat{\delta}_{GLS} &= \{Z'(I_G \otimes X)[\Sigma^{-1} \otimes (X'X)^{-1}](I_G \otimes X')Z\}^{-1} \\ &\quad \{Z'(I_G \otimes X)[\Sigma^{-1} \otimes (X'X)^{-1}](I_G \otimes X')y \end{aligned} \tag{11.44}$$

which upon simplifying yields

$$\widehat{\delta}_{GLS} = \{Z'[\Sigma^{-1} \otimes P_X]Z\}^{-1}\{Z'[\Sigma^{-1} \otimes P_X]y\} \tag{11.45}$$

Σ has to be estimated to make this estimator operational. Zellner and Theil (1962), suggest getting the 2SLS residuals for the i-th equation, say $\widehat{u}_i = y_i - Z_i \widehat{\delta}_{i,2SLS}$ and estimating Σ by $\widehat{\Sigma} = [\widehat{\sigma}_{ij}]$ where

$$\widehat{\sigma}_{ij} = [\widehat{u}_i' \widehat{u}_j / (T - g_i - k_i)^{1/2}(T - g_j - k_j)^{1/2}] \quad \text{for} \quad i, j = 1, 2, \dots, G.$$

If $\widehat{\Sigma}$ is substituted for Σ in (11.45), the resulting estimator is called 3SLS:

$$\widehat{\delta}_{3SLS} = \{Z'[\widehat{\Sigma}^{-1} \otimes P_X]Z\}^{-1}\{Z'[\widehat{\Sigma}^{-1} \otimes P_X]y\} \tag{11.46}$$

The asymptotic variance-covariance matrix of $\widehat{\delta}_{3SLS}$ can be estimated by $\{Z'[\widehat{\Sigma}^{-1} \otimes P_X]Z\}^{-1}$. If the system of equations (11.43) is properly specified, 3SLS is more efficient than 2SLS. But if say, the second equation is improperly specified while the first equation is properly specified, then a system estimator like 3SLS will be contaminated by this misspecification whereas a single equation estimator like 2SLS on the first equation is not. So, if the first equation is of interest it does not pay to go to a system estimator in this case.

Two sufficient conditions exist for the equivalence of 2SLS and 3SLS, these are the following: (i) Σ is diagonal, and (ii) every equation is just identified. Problem 5 leads you step by step through these results. It is also easy to show, see problem 5, that a necessary and sufficient condition for 3SLS to be equivalent to 2SLS on each equation is given by

$$\sigma^{ij} \widehat{Z}_i' \bar{P}_{\widehat{Z}_j} = 0 \quad \text{for} \quad i, j = 1, 2, \dots, G$$

where $\widehat{Z}_i = P_X Z_i$, see Baltagi (1989). This is similar to the condition derived in the seemingly unrelated regressions case except it involves the set of second stage regressors of 2SLS. One can easily see that besides the two sufficient conditions given above, $\widehat{Z}_i' \bar{P}_{\widehat{Z}_j} = 0$ states that the set of second stage regressors of the i-th equation have to be a perfect linear combination of those in the j-th equation and vice versa. A similar condition was derived by Kapteyn and Fiebig (1981). If some equations in the system are over-identified while others are just-identified, the 3SLS estimates of the over-identified equations can be obtained by running 3SLS ignoring the just-identified equations. The 3SLS estimates of each just-identified equation differ from those of 2SLS by a vector which is a linear function of the 3SLS residuals of the over-identified equations, see Theil (1971) and problem 17.

11.4 Test for Over-Identification Restrictions

We emphasized *instrument relevance*, now we turn to *instrument exogeneity*. Under just-identification, one cannot statistically test instruments for exogeneity. This choice of exogenous instruments requires making an expert judgement based on knowledge of the empirical application. However, if the first structural equation is over-identified, i.e., the number of instruments ℓ is larger than the number of right hand side variables $(g_1 + k_1)$, then one can test these over-identifying restrictions. A likelihood ratio test for this over-identification condition based on maximum likelihood procedures was given by Anderson and Rubin (1950). This version of the test requires the computation of LIML. This was later modified by Basmann (1960) so that it could be based on the 2SLS procedure. Here we present a simpler alternative based on Davidson and MacKinnon (1993) and Hausman (1983). In essence, one is testing

$$H_o;\ y_1 = Z_1\delta_1 + u_1 \quad \text{versus} \quad H_1;\ y_1 = Z_1\delta_1 + W^*\gamma + u_1 \tag{11.47}$$

where $u_1 \sim \text{IID}(0, \sigma_{11} I_T)$. Let W be the matrix of instruments of full rank ℓ. Also, let W^* be a subset of instruments W, of dimension $(\ell - k_1 - g_1)$, that are linearly independent of $\widehat{Z}_1 = P_W Z_1$. In this case, the matrix $[\widehat{Z}_1, W^*]$ has full rank ℓ and therefore, spans the same space as W. A test for over-identification is a test for $\gamma = 0$. In other words, W^* has no ability to explain any variation in y_1 that is not explained by Z_1 using the matrix of instruments W.

If W^* is correlated with u_1 or the first structural equation (11.34) is misspecified, say by Z_1 not including some variables in W^*, then $\gamma \neq 0$. Hence, testing $\gamma = 0$ should be interpreted as a joint test for the validity of the matrix of instruments W *and* the proper specification of (11.34) see, Davidson and MacKinnon (1993). Testing $H_o;\ \gamma = 0$ can be obtained as an asymptotic F-test as follows:

$$\frac{(RRSS^* - URSS^*)/\ell - (g_1 + k_1)}{URSS/(T - \ell)} \tag{11.48}$$

This is asymptotically distributed as $F(\ell - (g_1 + k_1), T - \ell)$ under H_o. Using instruments W, we regress Z_1 on W and get $\widehat{Z}_1$, then obtain the restricted 2SLS estimate $\widetilde{\delta}_{1,2SLS}$ by regressing y_1 on $\widehat{Z}_1$. The restricted residual sum of squares from the *second stage regression* is $RRSS^* = (y_1 - \widehat{Z}_1\widetilde{\delta}_{1,2SLS})'(y_1 - \widehat{Z}_1\widetilde{\delta}_{1,2SLS})$. Next we regress y_1 on $\widehat{Z}_1$ and W^* to get the unrestricted 2SLS estimates $\widehat{\delta}_{1,2SLS}$ and $\widehat{\gamma}_{2SLS}$. The unrestricted residual sum of squares from the *second stage regression* is $URSS^* = (y_1 - \widehat{Z}_1\widehat{\delta}_{1,2SLS} - W^*\widehat{\gamma}_{2SLS})'(y_1 - \widehat{Z}_1\widehat{\delta}_{1,2SLS} - W^*\widehat{\gamma}_{2SLS})$. The URSS

in (11.48) is the 2SLS residuals sum of squares from the unrestricted model which is obtained as follows $(y_1 - Z_1\widehat{\delta}_{1,2SLS} - W^*\widehat{\gamma}_{2SLS})'(y_1 - Z_1\widehat{\delta}_{1,2SLS} - W^*\widehat{\gamma}_{2SLS})$. URSS differs from $URSS^*$ in that Z_1 rather than $\widehat{Z}_1$ is used in obtaining the residuals. Note that this differs from the Chow-test in that the denominator is not based on $URSS^*$, see Wooldridge (1990).

This test does not require the construction of W^* for its implementation. This is because the model under H_1 is just-identified with as many regressor as there are instruments. This means that its

$$URSS^* = y_1'\bar{P}_W y_1 = y_1'y_1 - y_1'P_W y_1$$

see problem 10. It is easy to show, see problem 12, that

$$RRSS^* = y_1'\bar{P}_{\widehat{Z}_1} y_1 = y_1'y_1 - y_1'P_{\widehat{Z}_1} y_1$$

where $\widehat{Z}_1 = P_W Z_1$. Hence,

$$RRSS^* - URSS^* = y_1'P_W y_1 - y_1'P_{\widehat{Z}_1} y_1 \tag{11.49}$$

The test for over-identification can therefore be based on $RRSS^* - URSS^*$ divided by a *consistent* estimate of σ_{11}, say,

$$\widetilde{\sigma}_{11} = (y_1 - Z_1\widetilde{\delta}_{1,2SLS})'(y_1 - Z_1\widetilde{\delta}_{1,2SLS})/T \tag{11.50}$$

Problem 12 shows that the resulting test statistic is exactly that proposed by Hausman (1983). In a nutshell, the Hausman over-identification test regresses the 2SLS residuals $y_1 - Z_1\widetilde{\delta}_{1,2SLS}$ on the matrix W of *all* pre-determined variables in the model. The test statistic is T times the *uncentered* R^2 of this regression. See the Appendix to Chapter 3 for a definition of *uncentered* R^2. This test statistic is asymptotically distributed as χ^2 with $\ell - (g_1 + k_1)$ degrees of freedom. Large values of this statistic reject the null hypothesis.

Alternatively, one can get this test statistic as a Gauss-Newton Regression (GNR) on the unrestricted model in (11.47). To see this, recall from section 8.4 that the GNR applies to a general nonlinear model $y_t = x_t(\beta) + u_t$. Using the set of instruments W, the GNR becomes

$$y - x(\widetilde{\beta}) = P_W X(\widetilde{\beta})b + \text{residuals}$$

where $\widetilde{\beta}$ denotes the restricted instrumental variable estimate of β under the null hypothesis and $X(\beta)$ is the matrix of derivatives with typical elements $X_{ij}(\beta) = \partial x_i(\beta)/\partial\beta_j$ for $j = 1, \ldots, k$. Thus, the only difference between this GNR and that in Chapter 8 is that the regressors are multiplied by P_W, see Davidson and MacKinnon (1993, p. 226). Therefore, the GNR for (11.47) yields

$$y_1 - Z_1\widetilde{\delta}_{1,2SLS} = \widehat{Z}_1 b_1 + W^* b_2 + \text{residuals} \tag{11.51}$$

since $P_W[Z_1, W^*] = [\widehat{Z}_1, W^*]$ and $\widetilde{\delta}_{1,2SLS}$ is the restricted estimator under H_o; $\gamma = 0$. But, $[\widehat{Z}_1, W^*]$ spans the same space as W, see problem 12. Hence, the GNR in (11.51) is equivalent to running the 2SLS residuals on W and computing T times the *uncentered* R^2 as described above. Once again, it is clear that W^* need not be constructed.

The basic intuition behind the test for over-identification restriction rests on the fact that one can compute several legitimate IV estimators if all these instruments are relevant and

exogenous. For example, suppose there are two instruments and one right hand side endogenous variable. Then one can compute two IV estimators using each instrument separately. If these IV estimators produce very different estimates, then may be one instrument or the other or both are not exogenous. The over-identification test we just described implicitly makes this comparison without actually computing all possible IV estimates. Exogenous instruments have to be uncorrelated with the disturbances. This suggests that the 2SLS residuals have to be uncorrelated with the instruments. This is the basis for the TR_u^2 test statistic. If all the instruments are exogenous, the regression coefficient estimates should all be not significantly different from zero and the R_u^2 should be low.

11.5 Hausman's Specification Test[3]

A critical assumption for the linear regression model $y = X\beta + u$ is that the set of regressors X are uncorrelated with the error term u. Otherwise, we have simultaneous bias and OLS is inconsistent. Hausman (1978) proposed a general specification test for H_o; $E(u/X) = 0$ versus H_1; $E(u/X) \neq 0$. Two estimators are needed to implement this test. The first estimator must be a consistent and efficient estimator of β under H_o which becomes inconsistent under H_1. Let us denote this efficient estimator under H_o by $\widehat{\beta}_o$. The second estimator, denoted by $\widehat{\beta}_1$, must be consistent for β under both H_o and H_1, but inefficient under H_o. Hausman's test is based on the difference between these two estimators $\widehat{q} = \widehat{\beta}_1 - \widehat{\beta}_o$. Under H_o; plim $\widehat{q}$ is zero, while under H_1; plim $\widehat{q} \neq 0$. Hausman (1978) shows that $\text{var}(\widehat{q}) = \text{var}(\widehat{\beta}_1) - \text{var}(\widehat{\beta}_o)$ and Hausman's test becomes

$$m = \widehat{q}'[\text{var}(\widehat{q})]^{-1}\widehat{q} \tag{11.52}$$

which is asymptotically distributed under H_o as χ_k^2 where k is the dimension of β.

It remains to show that $\text{var}(\widehat{q})$ is the difference between the two variances. This can be illustrated for a single regressor case without matrix algebra, see Maddala (1992, page 507). First, one shows that $\text{cov}(\widehat{\beta}_o, \widehat{q}) = 0$. To prove this, consider a new estimator of β defined as $\widehat{\widehat{\beta}} = \widehat{\beta}_o + \lambda\widehat{q}$ where λ is an arbitrary constant. Under H_o, plim $\widehat{\widehat{\beta}} = \beta$ for every λ and

$$\text{var}(\widehat{\widehat{\beta}}) = \text{var}(\widehat{\beta}_o) + \lambda^2\text{var}(\widehat{q}) + 2\lambda\text{cov}(\widehat{\beta}_o, \widehat{q})$$

Since $\widehat{\beta}_o$ is efficient, $\text{var}(\widehat{\widehat{\beta}}) \geq \text{var}(\widehat{\beta}_o)$ which means that $\lambda^2\ \text{var}(\widehat{q}) + 2\lambda\ \text{cov}(\widehat{\beta}_o, \widehat{q}) \geq 0$ for every λ. If $\text{cov}(\widehat{\beta}_o, \widehat{q}) > 0$, then for $\lambda = -\text{cov}(\widehat{\beta}_o, \widehat{q})/\text{var}(\widehat{q})$ the above inequality is violated. Similarly, if $\text{cov}(\widehat{\beta}_o, \widehat{q}) < 0$, then for λ=-$\text{cov}(\widehat{\beta}_o, \widehat{q})/\text{var}(\widehat{q})$ the above inequality is violated. Therefore, under H_o, for the above inequality to be satisfied for every λ, it must be the case that $\text{cov}(\widehat{\beta}_o, \widehat{q}) = 0$.

Now, $\widehat{q} = \widehat{\beta}_1 - \widehat{\beta}_o$ can be rewritten as $\widehat{\beta}_1 = \widehat{q} + \widehat{\beta}_o$ with $\text{var}(\widehat{\beta}_1) = \text{var}(\widehat{q}) + \text{var}(\widehat{\beta}_o) + 2\text{cov}(\widehat{q}, \widehat{\beta}_o)$. Using the fact that the last term is zero, we get the required result: $\text{var}(\widehat{q}) = \text{var}(\widehat{\beta}_1) - \text{var}(\widehat{\beta}_o)$.

Example 4: Consider the simple regression without a constant

$$y_t = \beta x_t + u_t \qquad t = 1, 2, \ldots, T \quad \text{with} \quad u_t \sim \text{IIN}(0, \sigma^2)$$

and where β is a scalar. Under H_o; $E(u_t/x_t) = 0$ and OLS is efficient and consistent. Under H_1; OLS is not consistent. Using w_t as an instrumental variable, with w_t uncorrelated with u_t

and preferably highly correlated with x_t, yields the following IV estimator of β:

$$\widehat{\beta}_{IV} = \sum_{t=1}^{T} y_t w_t / \sum_{t=1}^{T} x_t w_t = \beta + \sum_{t=1}^{T} w_t u_t / \sum_{t=1}^{T} x_t w_t$$

with plim $\widehat{\beta}_{IV} = \beta$ under H_o and H_1 and $\text{var}(\widehat{\beta}_{IV}) = \sigma^2 \sum_{t=1}^{T} w_t^2 / (\sum_{t=1}^{T} x_t w_t)^2$. Also,

$$\widehat{\beta}_{OLS} = \sum_{t=1}^{T} x_t y_t / \sum_{t=1}^{T} x_t^2 = \beta + \sum_{t=1}^{T} x_t u_t / \sum_{t=1}^{T} x_t^2$$

with plim $\widehat{\beta}_{OLS} = \beta$ under H_o but plim $\widehat{\beta}_{OLS}$ under H_1, with $\text{var}(\widehat{\beta}_{OLS}) = \sigma^2 / \sum_{t=1}^{T} x_t^2$. In this case, $\widehat{q} = \widehat{\beta}_{IV} - \widehat{\beta}_{OLS}$ and plim $\widehat{q} \neq 0$ under H_1, while plim $\widehat{q} = 0$ under H_o, with

$$\begin{aligned} \text{var}(\widehat{q}) &= \text{var}(\widehat{\beta}_{IV}) - \text{var}(\widehat{\beta}_{OLS}) = \sigma^2 \left[\frac{\sum_{t=1}^{T} w_t^2}{\left(\sum_{t=1}^{T} x_t\right)^2} - \frac{1}{\sum_{t=1}^{T} x_t^2} \right] \\ &= \frac{\sigma^2}{\sum_{t=1}^{T} x_t^2} \left[\frac{1}{r_{xw}^2} - 1 \right] = \text{var}(\widehat{\beta}_{OLS}) \left[\frac{1 - r_{xw}^2}{r_{xw}^2} \right] \end{aligned}$$

where $r_{xw}^2 = (\sum_{t=1}^{T} x_t w_t)^2 / \sum_{t=1}^{T} w_t^2 \sum_{t=1}^{T} x_t^2$. Therefore, Hausman's test statistic is

$$m = \widehat{q}^2 r_{xw}^2 / \text{var}(\widehat{\beta}_{OLS})(1 - r_{xw}^2)$$

which is asymptotically distributed as χ_1^2 under H_o. Note that the same estimator of σ^2 is used for $\text{var}(\widehat{\beta}_{IV})$ and $\text{var}(\widehat{\beta}_{OLS})$. This is the estimator of σ^2 obtained under H_o.

The Hausman-test can also be obtained from the following *augmented regression*:

$$y_t = \beta x_t + \gamma \widehat{x}_t + \epsilon_t$$

where $\widehat{x}_t$ is the predicted value of x_t from regressing it on the instrumental variable w_t. Problem 13 asks the reader to show that Hausman's test statistic can be obtained by testing $\gamma = 0$.

In matrix form, the Durbin-Wu-Hausman test for the first structural equation is based upon the difference between OLS and IV estimation of (11.34) using the matrix of instruments W. In particular, the vector of contrasts is given by

$$\begin{aligned} \widehat{q} &= \widehat{\delta}_{1,IV} - \widehat{\delta}_{1,OLS} = (Z_1' P_W Z_1)^{-1} [Z_1' P_W y_1 - (Z_1' P_W Z_1)(Z_1' Z_1)^{-1} Z_1' y_1] \\ &= (Z_1' P_W Z_1)^{-1} [Z_1' P_W \bar{P}_{Z_1} y_1] \end{aligned} \tag{11.53}$$

Under the null hypothesis, $\widehat{q} = (Z_1' P_W Z_1)^{-1} Z_1' P_W \bar{P}_{Z_1} u_1$. The test for $\widehat{q} = 0$ can be based on the test for $Z_1' P_W \bar{P}_{Z_1} u_1$ having mean zero asymptotically. This last vector is of dimension $(g_1 + k_1)$. However, not all of its elements are necessarily random variables since $\bar{P}_{Z_1}$ may annihilate some columns of the second stage regressors $\widehat{Z}_1 = P_W Z_1$. In fact, all the included X's which are part of W, i.e., X_1, will be annihilated by $\bar{P}_{Z_1}$. Only the g_1 linearly independent variables $\hat{Y}_1 = P_W Y_1$ are not annihilated by $\bar{P}_{Z_1}$.

Our test focuses on the vector $\hat{Y}_1' \bar{P}_{Z_1} u_1$ having mean zero asymptotically. Now consider the artificial regression

$$y_1 = Z_1 \delta_1 + \hat{Y}_1 \gamma + \text{residuals} \tag{11.54}$$

Since $[Z_1, \hat{Y}_1]$, $[Z_1, \widehat{Z}_1]$, $[Z_1, Z_1 - \widehat{Z}_1]$ and $[Z_1, Y_1 - \widehat{Y}_1]$ all span the same column space, this regression has the same sum of squares residuals as

$$y_1 = Z_1 \delta_1 + (Y_1 - \hat{Y}_1) \eta + \text{residuals} \tag{11.55}$$

The DWH-test may be based on either of these regressions. It is equivalent to testing $\gamma = 0$ in (11.54) or $\eta = 0$ in (11.55) using an F-test. This is asymptotically distributed as $F(g_1, T - 2g_1 - k_1)$. Davidson and MacKinnon (1993, p. 239) warn about interpreting this test as one of exogeneity of Y_1 (the variables in Z_1 not in the space spanned by W). They argue that what is being tested is the consistency of the OLS estimates of δ_1, not that every column of Z_1 is independent of u_1.

In practice, one may be sure about using W_2 as a set of IV's but is not sure whether some r additional variables in Z_1 are legitimate as instruments. The DWH-test in this case will be based upon the difference between two IV estimators for δ_1. The first is $\widehat{\delta}_{2,IV}$ based on W_2 and the second is $\widehat{\delta}_{1,IV}$ based on W_1. The latter set includes W_2 and the additional r variables in Z_1.

$$\begin{aligned} \widehat{\delta}_{2,IV} - \widehat{\delta}_{1,IV} &= (Z_1'P_{W_2}Z_1)^{-1}[Z_1'P_{W_2}y_1 - (Z_1'P_{W_2}Z_1)(Z_1'P_{W_1}Z_1)^{-1}Z_1'P_{W_1}y_1] \\ &= (Z_1'P_{W_2}Z_1)^{-1}Z_1'P_{W_2}(\bar{P}_{P_{W_1}Z_1})y_1 \end{aligned} \tag{11.56}$$

since $P_{W_2}P_{W_1} = P_{W_2}$. The DWH-test is based on this contrast having mean zero asymptotically. Once again this last vector has dimension $g_1 + k_1$ and not all its elements are necessarily random variables since $\bar{P}_{P_{W_1}Z_1}$ annihilates some columns of $P_{W_2}Z_1$. This test can be based on the following artificial regression:

$$y_1 = Z_1\delta_1 + P_{W_2}Z_1^*\gamma + \text{residuals} \tag{11.57}$$

where $P_{W_2}Z_1^*$ consists of the r columns of $P_{W_2}Z_1$ that are not annihilated by $\bar{P}_{W_1}$. Regression (11.57) is performed with W_1 as the set of IV's and $\gamma = 0$ is tested using an F-test.

11.6 Empirical Example: Crime in North Carolina

Cornwell and Trumbull (1994) estimated an economic model of crime using data on 90 counties in North Carolina observed over the years 1981-87. This data set is available on the Springer web site as CRIME.DAT. Here, we consider cross-section data for 1987 and reconsider the full panel data set in Chapter 12. Table 11.2 gives the OLS estimates relating the crime rate (which is an FBI index measuring the number of crimes divided by the county population) to a set of explanatory variables. This was done using STATA. All variables are in logs except for the regional dummies. The explanatory variables consist of the probability of arrest (which is measured by ratio of arrests to offenses), probability of conviction given arrest (which is measured by the ratio of convictions to arrests), probability of a prison sentence given a conviction (measured by the proportion of total convictions resulting in prison sentences); average prison sentence in days as a proxy for sanction severity. The number of police per capita as a measure of the county's ability to detect crime, the population density which is the county population divided by county land area, a dummy variable indicating whether the county is in the SMSA with population larger than 50,000. Percent minority, which is the proportion of the county's population that is minority or non-white. Percent young male which is the proportion of the county's population that is males and between the ages of 15 and 24. Regional dummies for western and central counties. Opportunities in the legal sector captured by the average weekly wage in the county by industry. These industries are: construction; transportation, utilities and communication; wholesale and retail trade; finance, insurance and real estate; services; manufacturing; and federal, state and local government.

Results show that the probability of arrest as well as conviction given arrest have a negative and significant effect on the crime rate with estimated elasticities of −0.45 and −0.30 respectively. The probability of imprisonment given conviction as well as the sentence severity have a negative but insignificant effect on the crime rate. The greater the number of police per capita, the greater the number of reported crimes per capita. The estimated elasticity is 0.36 and it is significant. This could be explained by the fact that the larger the police force, the larger the reported crime. Alternatively, this could be an endogeneity problem with more crime resulting in the hiring of more police. The higher the density of the population the higher the crime rate. The estimated elasticity is 0.31 and it is significant. Returns to legal activity are insignificant except for wages in the service sector. This has a negative and significant effect on crime with an estimated elasticity of −0.27. Percent young male is insignificant, while percent minority is positive and significant with an estimated elasticity of 0.22. The central dummy variable is negative and significant while the western dummy variable is not significant. Also, the urban dummy variable is insignificant. Cornwell and Trumbull (1994) worried about the endogeneity of police per capita and the probability of arrest. They used as instruments two additional vari-

Table 11.2 Least Squares Estimates: Crime in North Carolina

Source	SS	df	MS		Number of obs =	90
					F(20,69) =	19.71
Model	22.8072483	20	1.14036241		Prob > F =	0.0000
Residual	3.99245334	69	.057861643		R-squared =	0.8510
					Adj R-squared =	0.8078
Total	26.7997016	89	.301120243		Root MSE =	.24054
lcrmrte	Coef.	Std. Err.	t	$P > \|t\|$	[95% Conf.	Interval]
lprbarr	−.4522907	.0816261	−5.54	0.000	−.6151303	−.2894511
lprbconv	−.3003044	.0600259	−5.00	0.000	−.4200527	−.180556
lprbpris	−.0340435	.1251096	−0.27	0.786	−.2836303	.2155433
lavgsen	−.2134467	.1167513	−1.83	0.072	−.4463592	.0194659
lpolpc	.3610463	.0909534	3.97	0.000	.1795993	.5424934
ldensity	.3149706	.0698265	4.51	0.000	.1756705	.4542707
lwcon	.2727634	.2198714	1.24	0.219	−.165868	.7113949
lwtuc	.1603777	.1666014	0.96	0.339	−.171983	.4927385
lwtrd	.1325719	.3005086	0.44	0.660	−.4669263	.7320702
lwfir	−.3205858	.251185	−1.28	0.206	−.8216861	.1805146
lwser	−.2694193	.1039842	−2.59	0.012	−.4768622	−.0619765
lwmfg	.1029571	.1524804	0.68	0.502	−.2012331	.4071472
lwfed	.3856593	.3215442	1.20	0.234	−.2558039	1.027123
lwsta	−.078239	.2701264	−0.29	0.773	−.6171264	.4606485
lwloc	−.1774064	.4251793	−0.42	0.678	−1.025616	.670803
lpctymle	.0326912	.1580377	0.21	0.837	−.2825855	.3479678
lpctmin	.2245975	.0519005	4.33	0.000	.1210589	.3281361
west	−.087998	.1243235	−0.71	0.481	−.3360167	.1600207
central	−.1771378	.0739535	−2.40	0.019	−.3246709	−.0296046
urban	−.0896129	.1375084	−0.65	0.517	−.3639347	.184709
_cons	−3.395919	3.020674	−1.12	0.265	−9.421998	2.630159

Table 11.3 Instrumental variables (2SLS) regression: Crime in North Carolina

Source	SS	df	MS		Number of obs =	90
					F(20,69) =	17.35
Model	22.6350465	20	1.13175232		Prob > F =	0.0000
Residual	4.16465515	69	.060357321		R-squared =	0.8446
					Adj R-squared =	0.7996
Total	26.7997016	89	.301120243		Root MSE =	.24568

lcrmrte	Coef.	Std. Err.	t	$P>\|t\|$	[95% Conf.	Interval]
lprbarr	−.4393081	.2267579	−1.94	0.057	−.8916777	.0130615
lpolpc	.5136133	.1976888	2.60	0.011	.1192349	.9079918
lprbconv	−.2713278	.0847024	−3.20	0.002	−.4403044	−.1023512
lprbpris	−.0278416	.1283276	−0.22	0.829	−.2838482	.2281651
lavgsen	−.280122	.1387228	−2.02	0.047	−.5568663	−.0033776
ldensity	.3273521	.0893292	3.66	0.000	.1491452	.505559
lwcon	.3456183	.2419206	1.43	0.158	−.137	.8282366
lwtuc	.1773533	.1718849	1.03	0.306	−.1655477	.5202542
lwtrd	.212578	.3239984	0.66	0.514	−.433781	.8589371
lwfir	−.3540903	.2612516	−1.36	0.180	−.8752731	.1670925
lwser	−.2911556	.1122454	−2.59	0.012	−.5150789	−.0672322
lwmfg	.0642196	.1644108	0.39	0.697	−.263771	.3922102
lwfed	.2974661	.3425026	0.87	0.388	−.3858079	.9807402
lwsta	.0037846	.3102383	0.01	0.990	−.615124	.6226931
lwloc	−.4336541	.5166733	−0.84	0.404	−1.464389	.597081
lpctymle	.0095115	.1869867	0.05	0.960	−.3635166	.3825397
lpctmin	.2285766	.0543079	4.21	0.000	.1202354	.3369179
west	−.0952899	.1301449	−0.73	0.467	−.3549219	.1643422
central	−.1792662	.0762815	−2.35	0.022	−.3314437	−.0270888
urban	−.1139416	.143354	−0.79	0.429	−.3999251	.1720419
_cons	−1.159015	3.898202	−0.30	0.767	−8.935716	6.617686

Instrumented:	lprbarr lpolpc
Instruments:	lprbconv lprbpris lavgsen ldensity lwcon lwtuc lwtrd lwfir lwser lwmfg lwfed lwsta lwloc lpctymle lpctmin west central ltaxpc lmix

ables. Offense mix which is the ratio of crimes involving face to face contact (such as robbery, assault and rape) to those that do not. The rationale for using this variable is that arrest is facilitated by positive identification of the offender. The second instrument is per capita tax revenue. This is justified on the basis that counties with preferences for law enforcement will vote for higher taxes to fund a larger police force.

The 2SLS estimates are reported in Table 11.3. The probability of arrest has an estimated elasticity of −0.44 but now with a p-value of 0.057. The probability of conviction given arrest has an estimated elasticity of −0.27 still significant. The probability of imprisonment given conviction is still insignificant while the sentence severity is now negative and significant with an estimated elasticity of −0.28. Police per capita has a higher elasticity of 0.51 still significant. The remaining estimates are slightly affected. In fact, the Hausman test based on the difference between the OLS and 2SLS estimates is shown in Table 11.4. This is computed

Table 11.4 Hausman's Test: Crime in North Carolina

	Coefficients			
	(b) b2sls	(B) bols	(b–B) Difference	sqrt(diag(V_b–V_B)) S.E.
lprbarr	−.4393081	−.4522907	.0129826	.2115569
lpolpc	.5136133	.3610463	.152567	.1755231
lprbconv	−.2713278	−.3003044	.0289765	.0597611
lprbpris	−.0278416	−.0340435	.0062019	.0285582
lavgsen	−.280122	−.2134467	−.0666753	.0749208
ldensity	.3273521	.3149706	.0123815	.0557132
lwcon	.3456183	.2727634	.0728548	.1009065
lwtuc	.1773533	.1603777	.0169755	.0422893
lwtrd	.212578	.1325719	.0800061	.1211178
lwfir	−.3540903	−.3205858	−.0335045	.0718228
lwser	−.2911556	−.2694193	−.0217362	.0422646
lwmfg	.0642196	.1029571	−.0387375	.0614869
lwfed	.2974661	.3856593	−.0881932	.1179718
lwsta	.0037846	−.078239	.0820236	.1525764
lwloc	−.4336541	−.1774064	−.2562477	.293554
lpctymle	.0095115	.0326912	−.0231796	.0999404
lpctmin	.2285766	.2245975	.0039792	.0159902
west	−.0952899	−.087998	−.0072919	.0384885
central	−.1792662	−.1771378	−.0021284	.0187016
urban	−.1139416	−.0896129	−.0243287	.0405192

b = consistent under Ho and Ha; obtained from ivreg

B = inconsistent under Ha, efficient under Ho; obtained from regress

Test: Ho: difference in coefficients not systematic

chi2(20) = (b−B)'[(V_b−V_B)^(−1)](b−B)
= 0.87
Prob > chi2 = 1.0000

using STATA and it contrasts 20 slope coefficient estimates. The Hausman test statistic is 0.87 and is asymptotically distributed as χ^2_{20}. This is insignificant, and shows that the 2SLS and OLS estimates are not significantly different given this model specification and the specific choice of instruments. Note that this is a just-identified equation and one cannot test for over-identification.

Tables 11.5 and 11.6 give the first-stage regressions for police per capita and the probability of arrest. The R^2 of these regressions are 0.56 and 0.47, respectively. The F-statistics for the significance of all slope coefficients are 4.42 and 3.11, respectively. The additional instruments (offense mix and per capita tax revenue) are jointly significant in both regressions yielding F-statistics of 10.56 and 5.78 with p-values of 0.0001 and 0.0048, respectively. Although there are two right hand side endogenous regressors in the crime equation rather than one, the Stock and Watson 'rule of thumb' suggest that these instruments may be weak.

Table 11.5 First Stage Regression: Police per Capita

Source	SS	df	MS		Number of obs =	90
					F(20,69) =	4.42
Model	6.99830344	20	0.349915172		Prob > F =	0.0000
Residual	5.46683312	69	0.079229465		R-squared =	0.5614
					Adj R-squared =	0.4343
Total	12.4651366	89	0.140057714		Root MSE =	0.28148
lpolpc	Coef.	Std. Err.	t	$P > \|t\|$	[95% Conf.	Interval]
lmix	0.2177256	0.0733414	2.97	0.004	0.0714135	0.3640378
ltaxpc	0.5601989	0.1489398	3.76	0.000	0.2630721	0.8573258
lprbconv	0.0037744	0.0744622	0.05	0.960	−0.1447736	0.1523223
lprbpris	−0.0487064	0.1470085	−0.33	0.741	−0.3419802	0.2445675
lavgsen	0.3958972	0.1295763	3.06	0.003	0.1373996	0.6543948
ldensity	0.0201292	0.0798454	0.25	0.802	−0.1391581	0.1794165
lwcon	−0.5368469	0.2562641	−2.09	0.040	−1.04808	−0.025614
lwtuc	−0.0216638	0.1955598	−0.11	0.912	−0.411795	0.3684674
lwtrd	−0.4207274	0.3483584	−1.21	0.231	−1.115683	0.2742286
lwfir	0.0001257	0.2976009	0.00	1.000	−0.5935718	0.5938232
lwser	0.0973089	0.1272819	0.76	0.447	−0.1566116	0.3512293
lwmfg	0.1710295	0.1814396	0.94	0.349	−0.1909327	0.5329916
lwfed	0.8555422	0.3779595	2.26	0.027	0.1015338	1.609551
lwsta	−0.1118764	0.3181352	−0.35	0.726	−0.7465387	0.5227859
lwloc	1.375102	0.4676561	2.94	0.004	0.4421535	2.30805
lpctymle	0.4186939	0.1869473	2.24	0.028	0.0457442	0.7916436
lpctmin	−0.0517966	0.0619159	−0.84	0.406	−0.1753154	0.0717222
west	0.1458865	0.1490133	0.98	0.331	−0.151387	0.4431599
central	0.0477227	0.0877814	0.54	0.588	−0.1273964	0.2228419
urban	−0.1192027	0.1719407	−0.69	0.490	−0.4622151	0.2238097
_cons	−16.33148	3.221824	−5.07	0.000	−22.75884	−9.904113

Notes

1. A heteroskedasticity-robust statistic is recommended especially if y_2 has discrete characteristics.
2. Why 10? See the proof in Appendix 10.4 of Stock and Watson (2003).
3. This test is also known as the Durbin-Wu-Hausman test, following the work of Durbin (1954), Wu (1973) and Hausman (1978).

Problems

1. Show that the OLS estimator of δ in (11.14), which can be written as

   $$\widehat{\delta}_{OLS} = \sum_{t=1}^{T} p_t q_t / \sum_{t=1}^{T} p_t^2$$

   is not consistent for δ. **Hint:** Write $\widehat{\delta}_{OLS} = \delta + \sum_{t=1}^{T} p_t(u_{2t} - \bar{u}_2) / \sum_{t=1}^{T} p_t^2$, and use (11.18) to show that

   $$\text{plim}\ \widehat{\delta}_{OLS} = \delta + (\sigma_{12} - \sigma_{22})(\delta - \beta)/[\sigma_{11} + \sigma_{22} - 2\sigma_{12}].$$

Table 11.6 First Stage Regression: Probability of Arrest

Source	SS	df	MS		Number of obs =	90
					F(20,69) =	3.11
Model	6.84874028	20	0.342437014		Prob > F =	0.0002
Residual	7.59345096	69	0.110050014		R-squared =	0.4742
					Adj R-squared =	0.3218
Total	14.4421912	89	0.162271812		Root MSE =	0.33174
lcrmrte	Coef.	Std. Err.	t	$P > \|t\|$	[95% Conf.	Interval]
lmix	0.2682143	0.0864373	3.10	0.003	0.0957766	0.4406519
ltaxpc	−0.1938134	0.1755345	−1.10	0.273	−0.5439952	0.1563684
lprbconv	−0.1946392	0.0877581	−2.22	0.030	−0.3697119	−0.0195665
lprbpris	−0.0240173	0.1732583	−0.14	0.890	−0.3696581	0.3216236
lavgsen	0.1565061	0.1527134	1.02	0.309	−0.1481488	0.4611611
ldensity	−0.2211654	0.0941026	−2.35	0.022	−0.408895	−0.0334357
lwcon	−0.2024569	0.3020226	−0.67	0.505	−0.8049755	0.4000616
lwtuc	−0.0461931	0.230479	−0.20	0.842	−0.5059861	0.4135999
lwtrd	0.0494793	0.4105612	0.12	0.904	−0.769568	0.8685266
lwfir	0.050559	0.3507405	0.14	0.886	−0.6491491	0.7502671
lwser	0.0551851	0.1500094	0.37	0.714	−0.2440754	0.3544456
lwmfg	0.0550689	0.2138375	0.26	0.798	−0.3715252	0.481663
lwfed	0.2622408	0.4454479	0.59	0.558	−0.6264034	1.150885
lwsta	−0.4843599	0.3749414	−1.29	0.201	−1.232347	0.2636277
lwloc	0.7739819	0.5511607	1.40	0.165	−0.3255536	1.873517
lpctymle	−0.3373594	0.2203286	−1.53	0.130	−0.776903	0.1021842
lpctmin	−0.0096724	0.0729716	−0.13	0.895	−0.1552467	0.1359019
west	0.0701236	0.1756211	0.40	0.691	−0.280231	0.4204782
central	0.0112086	0.1034557	0.11	0.914	−0.1951798	0.217597
urban	−0.0150372	0.2026425	−0.07	0.941	−0.4192979	0.3892234
_cons	−4.319234	3.797113	−1.14	0.259	−11.89427	3.255798

2. Consider equation (11.30) and let X_3 and X_4 be the only two other exogenous variables in this system.

   (a) Show that a two-stage estimator which regresses y_2 on X_1, X_2 and X_3 to get $y_2 = \hat{y}_2 + \hat{v}_2$, and y_3 on X_1, X_2 and X_4 to get $y_3 = \hat{y}_3 + \hat{v}_3$, and then regresses y_1 on $\hat{y}_2$, $\hat{y}_3$ and X_1 and X_2 does not necessarily yield consistent estimators. **Hint:** Show that the composite error is $\epsilon_1 = (u_1 + \alpha_{12}\hat{v}_2 + \alpha_{13}\hat{v}_3)$ and $\sum_{t=1}^{T} \hat{\epsilon}_{1t}\hat{y}_{2t} \neq 0$, because $\sum_{t=1}^{T} \hat{y}_{2t}\hat{v}_{3t} \neq 0$. The latter does not hold because $\sum_{t=1}^{T} X_{3t}\hat{v}_{3t} \neq 0$. (This shows that if both y's are not regressed on the *same set* of X's, the resulting two stage regression estimates are not consistent).

   (b) Show that the two-stage estimator which regresses y_2 and y_3 on X_2, X_3 and X_4 to get $y_2 = \hat{y}_2 + \hat{v}_2$ and $y_3 = \hat{y}_3 + \hat{v}_3$ and then regresses y_1 on $\hat{y}_2$, $\hat{y}_3$ and X_1 and X_2 is not necessarily consistent. **Hint:** Show that the composite error term $\epsilon_1 = u_1 + \alpha_{12}\hat{v}_2 + \alpha_{13}\hat{v}_3$ does not satisfy $\sum_{t=1}^{T} \hat{\epsilon}_{1t}X_{1t} = 0$, since $\sum_{t=1}^{T} \hat{v}_{2t}X_{1t} \neq 0$ and $\sum_{t=1}^{T} \hat{v}_{3t}X_{1t} \neq 0$. (This shows that if one of the *included* X's is not included in the first stage regression, then the resulting two-stage regression estimates are not consistent).

3. If equation (11.34) is just-identified, then X_2 is of the same dimension as Y_1, i.e., both are $T \times g_1$. Hence, Z_1 is of the same dimension as X, both of dimension $T \times (g_1 + k_1)$. Therefore, $X'Z_1$ is a square nonsingular matrix of dimension $(g_1 + k_1)$. Hence, $(Z_1'X)^{-1}$ exists. Using this fact, show

that $\widehat{\delta}_{1,2SLS}$ given by (11.36) reduces to $(X'Z_1)^{-1}X'y_1$. This is exactly the IV estimator with $W = X$, given in (11.41). Note that this is only feasible if $X'Z_1$ is square and nonsingular.

4. Premultiply equation (11.34) by X' and show that the transformed disturbances $X'u_1 \sim (0, \sigma_{11}(X'X))$. Perform GLS on the resulting transformed equation and show that $\widehat{\delta}_{1,GLS}$ is $\widehat{\delta}_{1,2SLS}$, given by (11.36).

5. *The Equivalence of 3SLS and 2SLS.*

   (a) Show that $\widehat{\delta}_{3SLS}$ given in (11.46) reduces to $\widehat{\delta}_{2SLS}$, when (i) Σ is diagonal, or (ii) every equation in the system is just-identified. **Hint:** For (i); show that $\widehat{\Sigma}^{-1} \otimes P_X$ is block-diagonal with the i-th block consisting of $P_X/\widehat{\sigma}_{ii}$. Also, Z is block-diagonal, therefore, $\{Z'[\widehat{\Sigma}^{-1} \otimes P_X]Z\}^{-1}$ is block-diagonal with the i-th block consisting of $\widehat{\sigma}_{ii}(Z_i'P_XZ_i)^{-1}$. Similarly, computing $Z'[\widehat{\Sigma}^{-1} \otimes P_X]y$, one can show that the i-th element of $\widehat{\delta}_{3SLS}$ is $(Z_i'P_XZ_i)^{-1}Z_i'P_Xy_i = \widehat{\delta}_{i,2SLS}$. For (ii); show that $Z_i'X$ is square and nonsingular under just-identification. Therefore, $\widehat{\delta}_{i,2SLS} = (X'Z_i)^{-1}X'y_i$ from problem 3. Also, from (11.44), we get

   $$\begin{aligned}\widehat{\delta}_{3SLS} &= \{\text{diag}[Z_i'X](\widehat{\Sigma}^{-1} \otimes (X'X)^{-1})\text{diag}[X'Z_i]\}^{-1} \\ &\quad \{\text{diag}[Z_i'X](\widehat{\Sigma}^{-1} \otimes (X'X)^{-1})(I_G \otimes X')y\}.\end{aligned}$$

   Using the fact that $Z_i'X$ is square, one can show that $\widehat{\delta}_{i,3SLS} = (X'Z_i)^{-1}X'y_i$.

   (b) Premultiply the system of equations in (11.43) by $(I_G \otimes P_X)$ and let $y^* = (I_G \otimes P_X)y$, $Z^* = (I_G \otimes P_X)Z$ and $u^* = (I_G \otimes P_X)u$, then $y^* = Z^*\delta + u^*$.Show that OLS on this transformed model yields 2SLS on each equation in (11.43). Show that GLS on this model yields 3SLS (knowing the true Σ) given in (11.45). Note that $\text{var}(u^*) = \Sigma \otimes P_X$ and its generalized inverse is $\Sigma^{-1} \otimes P_X$. Use the Milliken and Albohali condition for the equivalence of OLS and GLS given in equation (9.7) of Chapter 9 to deduce that 3SLS is equivalent to 2SLS if $Z^{*\prime}(\Sigma^{-1} \otimes P_X)\bar{P}_{Z^*} = 0$. Show that this reduces to the following necessary and sufficient condition $\sigma^{ij}\widehat{Z}_i'\bar{P}_{\widehat{Z}_j} = 0$ for $i \neq j$, see Baltagi (1989). **Hint:** Use the fact that

   $$Z^* = \text{diag}[P_XZ_i] = \text{diag}[\widehat{Z}_i] \quad \text{and} \quad \bar{P}_{Z^*} = \text{diag}[\bar{P}_{\widehat{Z}_i}].$$

   Verify that the two sufficient conditions given in part (a) satisfy this necessary and sufficient condition.

6. Consider the following demand and supply equations:

   $$\begin{aligned}Q &= a - bP + u_1 \\ Q &= c + dP + eW + fL + u_2\end{aligned}$$

   where W denotes weather conditions affecting supply, and L denotes the supply of immigrant workers available at harvest time.

   (a) Write this system in the matrix form given by equation (A.1) in the Appendix.

   (b) What does the *order-condition* for identification say about these two equations?

   (c) Premultiply this system by a nonsingular matrix $F = [f_{ij}]$, for $i, j = 1, 2$. What restrictions must the matrix F satisfy if the transformed model is to satisfy the same restrictions of the original model? Show that the first row of F is in fact the first row of an identity matrix, but the second row of F is not the second row of an identity matrix. What do you conclude?

7. Answer the same questions in problem 6 for the following model:

$$\begin{aligned} Q &= a - bP + cY + dA + u_1 \\ Q &= e + fP + gW + hL + u_2 \end{aligned}$$

where Y is real income and A is real assets.

8. Consider example (A.1) in the Appendix. Recall, that system of equations (A.3) and (A.4) are just-identified.

   (a) Construct ϕ for the demand equation (A.3) and show that $A\phi = (0, -f)'$ which is of rank 1 as long as $f \neq 0$. Similarly, construct ϕ for the supply equation (A.4) and show that $A\phi = (-c, 0)'$ which is of rank 1 as long as $c \neq 0$.

   (b) Using equation (A.17), show how the structural parameters can be retrieved from the reduced form parameters. Derive the reduced form equations for this system and verify the above relationships relating the reduced form and structural form parameters.

9. Derive the reduced form equations for the model given in problem 6, and show that the structural parameters of the second equation cannot be derived from the reduced form parameters. Also, show that there are more than one way of expressing the structural parameters of the first equation in terms of the reduced form parameters.

10. *Just-Identified Model.* Consider the just-identified equation

$$y_1 = Z_1\delta_1 + u_1$$

with W, the matrix of instruments for this equation of dimension $T \times \ell$ where $\ell = g_1 + k_1$ the dimension of Z_1. In this case, $W'Z_1$ is square and nonsingular.

   (a) Show that the *generalized instrumental variable* estimator given below (11.41) reduces to the *simple instrumental variable* estimator given in (11.38).

   (b) Show that the minimized value of the criterion function for this just-identified model is zero, i.e., show that $(y_1 - Z_1\widehat{\delta}_{1,IV})'P_W(y_1 - Z_1\widehat{\delta}_{1,IV}) = 0$.

   (c) Conclude that the residual sum of squares of the second stage regression of this just-identified model is the same as that obtained by regressing y_1 on the matrix of instruments W, i.e., show that $(y_1 - \widehat{Z}_1\widehat{\delta}_{1,IV})'(y_1 - \widehat{Z}_1\widehat{\delta}_{1,IV}) = y_1'\bar{P}_W y_1$ where $\widehat{Z}_1 = P_W Z_1$. **Hint:** Show that $P_{\widehat{Z}_1} = P_{P_W Z_1} = P_W$, under just-identification.

11. Let W_1 and W_2 be two sets of instrumental variables for the first structural equation given in (11.34). Suppose that W_1 is spanned by the space of W_2. Verify that the resulting IV estimator of δ_1 based on W_2 is at least as efficient as that based on W_1. **Hint:** Show that $P_{W_2}W_1 = W_1$ and that $P_{W_2} - P_{W_1}$ is idempotent. Conclude that the difference in the corresponding asymptotic covariances of these IV estimators is positive semi-definite. (This shows that increasing the number of legitimate instruments should improve the asymptotic efficiency of an IV estimator).

12. *Testing for Over-Identification.* In testing H_o; $\gamma = 0$ versus H_1; $\gamma \neq 0$ in section 11.4, equation (11.47):

   (a) Show that the second stage regression of 2SLS on the unrestricted model $y_1 = Z_1\delta_1 + W^*\gamma + u_1$ with the matrix of instruments W yields the following residual sum of squares:

$$URSS^* = y_1'\bar{P}_W y_1 = y_1'y_1 - y_1'P_W y_1$$

   Hint: Use the results of problem 10 for the just-identified case.

(b) Show that the second stage regression of 2SLS on the restricted model $y_1 = Z_1\delta_1 + u_1$ with the matrix of instruments W yields the following residual sum of squares:

$$RRSS^* = y_1'\bar{P}_{\widehat{Z}_1}y_1 = y_1'y_1 - y_1'P_{\widehat{Z}_1}y_1$$

where $\widehat{Z}_1 = P_W Z_1$ and $P_{\widehat{Z}_1} = P_W Z_1(Z_1'P_W Z_1)^{-1}Z_1'P_W$. Conclude that $RRSS^* - URSS^*$ yields (11.49).

(c) Consider the test statistic $(RRSS^* - URSS^*)/\widehat{\sigma}_{11}$ where $\widehat{\sigma}_{11}$ is given by (11.50) as the usual 2SLS residual sum of squares under H_o divided by T. Show that it can be written as Hausman's (1983) test statistic, i.e., nR_u^2 where R_u^2 is the *uncentered* R^2 of the regression of 2SLS residuals $(y_1 - Z_1\widehat{\delta}_{1,2SLS})$ on the matrix of all pre-determined variables W. **Hint:** Show that the regression sum of squares $(y_1 - Z_1\widehat{\delta}_{1,2SLS})'P_W(y_1 - Z_1\widehat{\delta}_{1,2SLS}) = (RRSS^* - URSS^*)$ given in (11.49).

(d) Verify that the test for H_o based on the GNR for the model given in part (a) yields the same TR_u^2 test statistic described in part (c).

13. *Hausman's Specification Test:* OLS Versus 2SLS. This is based on Maddala (1992, page 511). For the simple regression

$$y_t = \beta x_t + u_t \quad t = 1, 2 \ldots, T$$

where β is scalar and $u_t \sim \text{IIN}(0, \sigma^2)$. Let w_t be an instrumental variable for x_t. Run x_t on w_t and get $x_t = \widehat{\pi}w_t + \widehat{\nu}_t$ or $x_t = \hat{x}_t + \widehat{\nu}_t$ where $\hat{x}_t = \widehat{\pi}w_t$.

(a) Show that in the *augmented regression* $y_t = \beta x_t + \gamma\hat{x}_t + \epsilon_t$ a test for $\gamma = 0$ based on OLS from this regression yields Hausman's test-statistic. **Hint:** Show that $\widehat{\gamma}_{OLS} = \widehat{q}/(1 - r_{xw}^2)$ where

$$r_{xw}^2 = \left(\sum_{t=1}^T x_t w_t\right)^2 / \sum_{t=1}^T w_t^2 \sum_{t=1}^T x_t^2.$$

Next, show that $\text{var}(\widehat{\gamma}_{OLS}) = \text{var}(\widehat{\beta}_{OLS})/r_{xw}^2(1 - r_{xw}^2)$. Conclude that

$$\widehat{\gamma}_{OLS}^2/\text{var}(\widehat{\gamma}_{OLS}) = \widehat{q}^2 r_{xw}^2/[\text{var}(\widehat{\beta}_{OLS})(1 - r_{xw}^2)]$$

is the Hausman (1978) test statistic m given in section 11.5.

(b) Show that the same result in part (a) could have been obtained from the *augmented regression*

$$y_t = \beta x_t + \gamma\widehat{\nu}_t + \eta_t$$

where $\widehat{\nu}_t$ is the residual from the regression of x_t on w_t.

14. Consider the following structural equation: $y_1 = \alpha_{12}y_2 + \alpha_{13}y_3 + \beta_{11}X_1 + \beta_{12}X_2 + u_1$ where y_2 and y_3 are endogenous and X_1 and X_2 are exogenous. Also, suppose that the excluded exogenous variables include X_3 and X_4.

(a) Test the null hypothesis that y_2 and y_3 are exogenous against the alternative that they are not. Show that Hausman's test statistic can be obtained from the augmented regression:

$$y_1 = \alpha_{12}y_2 + \alpha_{13}y_3 + \beta_{11}X_1 + \beta_{12}X_2 + \gamma_2\hat{y}_2 + \gamma_3\hat{y}_3 + \epsilon_1$$

where $\hat{y}_2$ and $\hat{y}_3$ are predicted values from regressing y_2 and y_3 on $X = [X_1, X_2, X_3, X_4]$. Hausman's test is equivalent to testing H_o; $\gamma_2 = \gamma_3 = 0$. See equation (11.54).

(b) Show that the same results in part (a) hold if we had used the following augmented regression:

$$y_1 = \alpha_{12}y_2 + \alpha_{13}y_3 + \beta_{11}X_1 + \beta_{12}X_2 + \gamma_2\widehat{\nu}_2 + \gamma_3\widehat{\nu}_3 + \eta_1$$

where $\widehat{\nu}_2$ and $\widehat{\nu}_3$ are the residuals from running y_2 and y_3 on $X = [X_1, X_2, X_3, X_4]$. See equation (11.55). **Hint:** Show that the regressions in (a) and (b) have the same residual sum of squares.

15. For the artificial regression given in (11.55):

(a) Show that OLS on this model yields $\widehat{\delta}_{1,OLS} = \widehat{\delta}_{1,IV} = (Z_1'P_WZ_1)^{-1}Z_1'P_Wy_1$. **Hint:** $Y_1 - \hat{Y}_1 = \bar{P}_WY_1$. Use the FWL Theorem to residual out these variables in (11.55) and use the fact that $\bar{P}_WZ_1 = [\bar{P}_WY_1, 0]$.

(b) Show that the $\text{var}(\widehat{\delta}_{1,OLS}) = \widetilde{s}_{11}(Z_1'P_WZ_1)^{-1}$ where $\widetilde{s}_{11}$ is the mean squared error of the OLS regression in (11.55). Note that when $\eta \neq 0$ in (11.55), IV estimation is necessary and $\widetilde{s}_{11}$ underestimates σ_{11} and will have to be replaced by $(y_1 - Z_1\widehat{\delta}_{1,IV})'(y_1 - Z_1\widehat{\delta}_{1,IV})/T$.

16. *Recursive Systems.* A recursive system has two crucial features: B is a triangular matrix and Σ is a diagonal matrix. For this special case of the simultaneous equations model, OLS is still consistent, and under normality of the disturbances still maximum likelihood. Let us consider a specific example:

$$\begin{aligned} y_{1t} + \gamma_{11}x_{1t} + \gamma_{12}x_{2t} &= u_{1t} \\ \beta_{21}y_{1t} + y_{2t} + \gamma_{23}x_{3t} &= u_{2t} \end{aligned}$$

In this case, $B = \begin{bmatrix} 1 & 0 \\ \beta_{21} & 1 \end{bmatrix}$ is triangular and $\sum = \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix}$ is assumed diagonal.

(a) Check the identifiability conditions of this recursive system.

(b) Solve for the reduced form and show that y_{1t} is only a function of the x_t's and u_{1t}, while y_{2t} is a function of the x_t's and a linear combination of u_{1t} and u_{2t}.

(c) Show that OLS on the first structural equation yields consistent estimates. **Hint:** There are no right hand side y's for the first equation. Show that despite the presence of y_1 in the second equation, OLS of y_2 on y_1 and x_3 yields consistent estimates. Note that y_1 is a function of u_1 only and u_1 and u_2 are not correlated.

(d) Under the normality assumption on the disturbances, the likelihood function conditional on the x's is given by

$$L(B, \Gamma, \Sigma) = (2\pi)^{-T/2}|B|^T|\Sigma|^{-T/2}\exp(-\frac{1}{2}\textstyle\sum_{t=1}^{T} u_t'\Sigma^{-1}u_t)$$

where in this two equation case $u_t' = (u_{1t}, u_{2t})$. Since B is triangular, $|B| = 1$. Show that maximizing L with respect to B and Γ is equivalent to minimizing $Q = \sum_{t=1}^{T} u_t'\Sigma^{-1}u_t$. Conclude that when Σ is diagonal, Σ^{-1} is diagonal and $Q = \sum_{t=1}^{T} u_{1t}^2/\sigma_{11} + \sum_{t=1}^{T} u_{2t}^2/\sigma_{22}$. Hence, maximizing the likelihood with respect to B and Γ is equivalent to running OLS on each equation *separately*.

17. *Hausman's Specification Test: 2SLS Versus 3SLS.* This is based on Holly (1988). Consider the two-equations model,

$$\begin{aligned} y_1 &= \alpha y_2 + \beta_1 x_1 + \beta_2 x_2 + u_1 \\ y_2 &= \gamma y_1 + \beta_3 x_3 + u_2 \end{aligned}$$

where y_1 and y_2 are endogenous; x_1, x_2 and x_3 are exogenous (the y's and the x's are $n \times 1$ vectors). The standard assumptions are made on the disturbance vectors u_1 and u_2. With the usual notation, the model can also be written as

$$\begin{aligned} y_1 &= Z_1\delta_1 + u_1 \\ y_2 &= Z_2\delta_2 + u_2 \end{aligned}$$

The following notation will be used: $\widetilde{\delta} = 2SLS$, $\widetilde{\widetilde{\delta}} = 3SLS$, and the corresponding residuals will be denoted as $\widetilde{u}$ and $\widetilde{\widetilde{u}}$, respectively.

(a) Assume that $\alpha\gamma \neq 1$. Show that the 3SLS estimating equations reduce to

$$\begin{aligned} &\widetilde{\sigma}^{11}X'\widetilde{\widetilde{u}}_1 + \widetilde{\sigma}^{12}X'\widetilde{\widetilde{u}}_2 = 0 \\ &\widetilde{\sigma}^{12}Z_2'P_X\widetilde{\widetilde{u}}_1 + \widetilde{\sigma}^{22}Z_2'P_X\widetilde{\widetilde{u}}_2 = 0 \end{aligned}$$

where $X = (x_1, x_2, x_3)$, $\Sigma = [\sigma_{ij}]$ is the structural form covariance matrix, and $\Sigma^{-1} = [\sigma^{ij}]$ for $i, j = 1, 2$.

(b) Deduce that $\widetilde{\widetilde{\delta}}_2 = \widetilde{\delta}_2$ and $\widetilde{\widetilde{\delta}}_1 = \widetilde{\delta}_1 - (\widetilde{\sigma}_{12}/\widetilde{\sigma}_{22})(Z_1'P_XZ_1)^{-1}Z_1'P_X\widetilde{u}_2$. This proves that the 3SLS estimator of the over-identified second equation is equal to its 2SLS counterpart. Also, the 3SLS estimator of the just-identified first equation differs from its 2SLS (or indirect least squares) counterpart by a linear combination of the 2SLS (or 3SLS) residuals of the over-identified equation, see Theil (1971).

(c) How would you interpret a Hausman-type test where you compare $\widetilde{\widetilde{\delta}}_1$ and $\widetilde{\delta}_1$? Show that it is nR^2 where R^2 is the R-squared of the regression of $\widetilde{u}_2$ on the set of second stage regressors of both equations $\widehat{Z}_1$ and $\widehat{Z}_2$. **Hint:** See the solution by Baltagi (1989).

18. For the two-equation simultaneous model

$$\begin{aligned} y_{1t} &= \beta_{12}y_{2t} + \gamma_{11}x_{1t} + u_{1t} \\ y_{2t} &= \beta_{21}y_{1t} + \gamma_{22}x_{2t} + \gamma_{23}x_{3t} + u_{2t} \end{aligned}$$

With

$$X'X = \begin{bmatrix} 20 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 10 \end{bmatrix} \quad X'Y = \begin{bmatrix} 5 & 10 \\ 40 & 20 \\ 20 & 30 \end{bmatrix} \quad Y'Y = \begin{bmatrix} 3 & 4 \\ 4 & 8 \end{bmatrix}$$

(a) Determine the identifiability of each equation with the aid of the order and rank conditions for identification.

(b) Obtain the OLS normal equations for both equations. Solve for the OLS estimates.

(c) Obtain the 2SLS normal equations for both equations. Solve for the 2SLS estimates.

(d) Can you estimate these equations using Indirect Least Squares? Explain.

19. Laffer (1970) considered the following supply and demand equations for Traded Money:

$$\begin{aligned} \log(TM/P) &= \alpha_o + \alpha_1\log(RM/P) + \alpha_2\log i + u_1 \\ \log(TM/P) &= \beta_o + \beta_1\log(Y/P) + \beta_2\log i + \beta_3\log(S1) + \beta_4\log(S2) + u_2 \end{aligned}$$

where

$$\begin{aligned} TM &= \text{Nominal total trade money} \\ RM &= \text{Nominal effective reserve money} \\ Y &= \text{GNP in current dollars} \\ S2 &= \text{Degree of market utilization} \\ i &= \text{short-term rate of interest} \\ S1 &= \text{Mean real size of the representative economic unit } (1939 = 100) \\ P &= \text{GNP price deflator } (1958 = 100) \end{aligned}$$

The basic idea is that trade credit is a line of credit and the unused portion represents purchasing power which can be used as a medium of exchange for goods and services. Hence, Laffer (1970) suggests that trade credit should be counted as part of the money supply. Besides real income and the short-term interest rate, the demand for real traded money includes $\log(S1)$ and $\log(S2)$. $S1$ is included to capture economies of scale. As $S1$ increases, holding everything else constant, the presence of economies of scale would mean that the demand for traded money would decrease. Also, the larger $S2$, the larger the degree of market utilization and the more money is needed for transaction purposes.

The data are provided on the Springer web site as LAFFER.ASC. This data covers 21 annual observations over the period 1946-1966. This was obtained from Lott and Ray (1992). Assume that (TM/P) and i are endogenous and the rest of the variables in this model are exogenous.

(a) Using the *order condition* for identification, determine whether the demand and supply equations are identified? What happens if you used the *rank condition* of identification?

(b) Estimate this model using OLS.

(c) Estimate this model using 2SLS.

(d) Estimate this model using 3SLS. Compare the estimates and their standard errors for parts (b), (c) and (d).

(e) Test the over-identification restriction of each equation.

(f) Run Hausman's specification test on each equation basing it on OLS and 2SLS.

(g) Run Hausman's specification test on each equation basing it on 2SLS and 3SLS.

20. The market for a certain good is expressed by the following equations:

$$\begin{aligned} D_t &= \alpha_0 - \alpha_1 P_t + \alpha_2 X_t + u_{1t} \qquad (\alpha_1, \alpha_2 > 0) \\ S_t &= \beta_0 + \beta_1 P_t + u_{2t} \qquad (\beta_1 > 0) \\ D_t &= S_t = Q_t \end{aligned}$$

where D_t is the quantity demanded, S_t is the quantity supplied, X_t is an exogenous demand shift variable. (u_{1t}, u_{2t}) is an IID random vector with zero mean and covariance matrix $\Sigma = [\sigma_{ij}]$ for $i, j = 1, 2$.

(a) Examine the identifiability of the model under the assumptions given above using the order and rank conditions of identification.

(b) Assuming the moment matrix of exogenous variables converge to a finite non-zero matrix, derive the simultaneous equation bias in the OLS estimator of β_1.

(c) If $\sigma_{12} = 0$ would you expect this bias to be positive or negative? Explain.

21. Consider the following three equations simultaneous model

$$\begin{aligned} y_1 &= \alpha_1 + \beta_2 y_2 + \gamma_1 X_1 + u_1 \quad (1)\\ y_2 &= \alpha_2 + \beta_1 y_1 + \beta_3 y_3 + \gamma_2 X_2 + u_2 \quad (2)\\ y_3 &= \alpha_3 + \gamma_3 X_3 + \gamma_4 X_4 + \gamma_5 X_5 + u_3 \quad (3) \end{aligned}$$

where the X's are exogenous and the y's are endogenous.

(a) Examine the identifiability of this system using the order and rank conditions.

(b) How would you estimate equation (2) by 2SLS? Describe your procedure step by step.

(c) Suppose that equation (1) was estimated by running y_2 on a constant X_2 and X_3 and the resulting predicted $\hat{y}_2$ was substituted in (1), and OLS performed on the resulting model. Would this estimating procedure yield consistent estimates of α_1, β_2 and γ_1? Explain your answer.

(d) How would you test for the over-identification restrictions in equation (1)?

22. *Equivariance of Instrumental Variables Estimators.* This is based on Sapra (1997). For the structural equation given in (11.34), let the matrix of instruments W be of dimension $T \times \ell$ where $\ell \geq g_1 + k_1$ as described below (11.41). Then the corresponding instrumental variable estimator of δ_1 given below (11.41) is $\widehat{\delta}_{1,IV}(y_1) = (Z_1' P_W Z_1)^{-1} Z_1' P_W y_1$.

(a) Show that this IV estimator is an equivariant estimator of δ_1, i.e., show that for any linear transformation $y_1^* = a y_1 + Z_1 b$ where a is a positive scalar and b is an $(\ell \times 1)$ real vector, the following relationship holds:

$$\widehat{\delta}_{1,IV}(y_1^*) = a\widehat{\delta}_{1,IV}(y_1) + b.$$

(b) Show that the variance estimator

$$\widehat{\sigma}^2(y_1) = (y_1 - Z_1\widehat{\delta}_{1,IV}(y_1))'(y_1 - Z_1\widehat{\delta}_{1,IV}(y_1))/T$$

is equivariant for σ^2, i.e., show that $\widehat{\sigma}^2(y_1^*) = a^2\widehat{\sigma}^2(y_1)$.

23. *Identification and Estimation of a Simple Two-Equation Model.* This is based on Holly (1987). Consider the following two equation model

$$\begin{aligned} y_{t1} &= \alpha + \beta y_{t2} + u_{t1}\\ y_{t2} &= \gamma + y_{t1} + u_{t2} \end{aligned}$$

where the y's are endogenous variables and the u's are serially independent disturbances that are identically distributed with zero means and nonsingular covariance matrix $\Sigma = [\sigma_{ij}]$ where $E(u_{ti}u_{tj}) = \sigma_{ij}$ for $i, j = 1, 2$, and all $t = 1, 2, \ldots, T$. The reduced form equations are given by

$$y_{t1} = \pi_{11} + \nu_{t1} \quad \text{and} \quad y_{t2} = \pi_{21} + \nu_{t2}$$

with $\Omega = [\omega_{ij}]$ where $E(\nu_{ti}\nu_{tj}) = \omega_{ij}$ for $i, j = 1, 2$ and all $t = 1, 2, \ldots, T$.

(a) Examine the identification of this system of two equations when no further information is available.

(b) Repeat part (a) when $\sigma_{12} = 0$.

(c) Assuming $\sigma_{12} = 0$, show that $\widehat{\beta}_{OLS}$, the OLS estimator of β in the first equation is not consistent.

(d) Assuming $\sigma_{12} = 0$, show that an alternative consistent estimator of β is an IV estimator using $z_t = [(y_{t2} - \bar{y}_2) - (y_{t1} - \bar{y}_1)]$ as an instrument for y_{t2}.

(e) Show that the IV estimator of β obtained from part (d) is also an indirect least squares estimator of β. **Hint:** See the solution by Singh and Bhat (1988).

24. *Errors in Measurement and the Wald (1940) Estimator.* This is based on Farebrother (1985). Let y_i^* be permanent consumption and X^* be permanent income, both are measured with error:

$$y_i^* = \beta x_i^* \quad \text{where} \quad y_i = y_i^* + \epsilon_i \quad \text{and} \quad x_i = x_i^* + u_i \quad \text{for} \quad i = 1, 2, \ldots, n.$$

Let x_i^*, ϵ_i and u_i be independent normal random variables with zero means and variances σ_*^2, σ_ϵ^2 and σ_u^2, respectively. Wald (1940) suggested the following estimator of β: Order the sample by the x_i's and split the sample into two. Let $(\bar{y}_1, \bar{x}_1)$ be the sample mean of the first half of the sample and $(\bar{y}_2, \bar{x}_2)$ be the sample mean of the second half of this sample. Wald's estimator of β is $\widehat{\beta}_W = (\bar{y}_2 - \bar{y}_1)/(\bar{x}_2 - \bar{x}_1)$. It is the slope of the line joining these two sample mean observations.

(a) Show that $\widehat{\beta}_W$ can be interpreted as a simple IV estimator with instrument

$$\begin{aligned} z_i &= 1 \quad \text{for} \quad x_i \geq \text{median}(x) \\ &= -1 \quad \text{for} \quad x_i < \text{median}(x) \end{aligned}$$

where median(x) is the sample median of $x_1, x_2, \ldots, x_n$.

(b) Define $w_i = \rho^2 x_i^* - \tau^2 u_i$ where $\rho^2 = \sigma_u^2/(\sigma_u^2 + \sigma_*^2)$ and $\tau^2 = \sigma_*^2/(\sigma_u^2 + \sigma_*^2)$. Show that $E(x_i w_i) = 0$ and that $w_i \sim N(0, \sigma_*^2 \sigma_u^2/(\sigma_*^2 + \sigma_u^2))$.

(c) Show that $x_i^* = \tau^2 x_i + w_i$ and use it to show that

$$E(\widehat{\beta}_W / x_1, \ldots, x_n) = E(\widehat{\beta}_{OLS} / x_1, \ldots, x_n) = \beta \tau^2.$$

Conclude that the exact small sample bias of $\widehat{\beta}_{OLS}$ and $\widehat{\beta}_W$ are the same.

25. *Comparison of t-ratios.* This is based on Holly (1990). Consider the two equations model

$$y_1 = \alpha y_2 + X\beta + u_1 \quad \text{and} \quad y_2 = \gamma y_1 + X\beta + u_2$$

where α and γ are scalars, y_1 and y_2 are $T \times 1$ and X is a $T \times (K-1)$ matrix of exogenous variables. Assume that $u_i \sim N(0, \sigma_i^2 I_T)$ for $i = 1, 2$. Show that the t-ratios for H_o^a; $\alpha = 0$ and H_o^b; $\gamma = 0$ using $\widehat{\alpha}_{OLS}$ and $\widehat{\gamma}_{OLS}$ are the same. Comment on this result. **Hint:** See the solution by Farebrother (1991).

26. *Degeneration of Feasible GLS to 2SLS in a Limited Information Simultaneous Equations Model.* This is based on Gao and Lahiri (2000). Consider a simple limited information simultaneous equations model,

$$y_1 = \gamma y_2 + u, \tag{1}$$
$$y_2 = X\beta + v, \tag{2}$$

where y_1, y_2 are $N \times 1$ vectors of observations on two endogenous variables. X is $N \times K$ matrix of predetermined variables of the system, and $K \geq 1$ such that (1) is identified. Each row of (u, v) is assumed to be i.i.d. $(0, \Sigma)$, and Σ is p.d. In this case, $\widehat{\gamma}_{2SLS} = (y_2' P_X y_2)^{-1} y_2' P_X y_1$, where $P_X = X(X'X)^{-1}X'$. The residuals $\widehat{u} = y_1 - \widehat{\gamma}_{2SLS} y_2$ and $\widehat{v} = M y_2$, where $M = I_N - P_X$ are used to generate a consistent estimate for Σ

$$\widehat{\Sigma} = \frac{1}{N} \begin{bmatrix} \widehat{u}'\widehat{u} & \widehat{u}'\widehat{v} \\ \widehat{v}'\widehat{u} & \widehat{v}'\widehat{v} \end{bmatrix}.$$

Show that a feasible GLS estimate of γ using $\widehat{\Sigma}$ degenerates to $\widehat{\gamma}_{2SLS}$.

27. *Equality of Two IV Estimators.* This is based on Qian (1999). Consider the following linear regression model:

$$y_i = x_i'\beta + \epsilon = x_{1i}'\beta_1 + x_{2i}'\beta_2 + \epsilon_i, \qquad i = 1, 2, \ldots, N, \tag{1}$$

where the dimensions of x_i', x_{1i}' and x_{2i}' are $1\times K$, $1\times K_1$ and $1\times K_2$, respectively, with $K = K_1+K_2$. x_i may be correlated with ϵ_i, but we have instruments z_i such that $E(\epsilon_i|z_i') = 0$ and $E(\epsilon_i^2|z_i') = \sigma^2$. Partition the instruments into two subsets: $z_i' = (z_{1i}', z_{2i}')$, where the dimensions of z_i, z_{1i}, and z_{2i} are L, L_1 and L_2, with $L = L_1 + L_2$. Assume that $E(z_{1i}x_i')$ has full column rank (so $L_1 \geq K$); this ensures that β can be estimated consistently using the subset of instruments z_{1i} only or using the entire set $z_i' = (z_{1i}', z_{2i}')$. We also assume that (y_i, x_i', z_i') is covariance stationary.

Define $P_A = A(A'A)^{-1}A'$ and $M_A = I - P_A$ for any matrix A with full column rank. Let $X = (x_1, \ldots, x_N)'$, and similarly for X_1, X_2, y, Z_1, Z_2 and Z. Define $\widehat{X} = P_{[Z_1]}X$ and $\widehat{\beta} = (\widehat{X}'\widehat{X})^{-1}\widehat{X}'y$, so that $\widehat{\beta}$ is the instrumental variables (IV) estimator of (1) using Z_1 as instruments. Similarly, define $\widetilde{X} = P_Z X$ and $\widetilde{\beta} = (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'y$, so that $\widetilde{\beta}$ is the IV estimator of (1) using Z as instruments.

Show that $\widehat{\beta}_1 = \widetilde{\beta}_1$ if $Z_2'M_{[Z_1]}[X_1 - X_2(X_2'P_1X_2)^{-1}X_2'P_1X_1] = 0$.

28. For the crime in North Carolina example given in section 11.6, replicate the results in Tables 11.2-11.6 using the data for 1987. Do the same using the data for 1981. Are there any notable differences in the results as we compare 1981 to 1987?

References

This chapter is influenced by Johnston (1984), Kelejian and Oates (1989), Maddala (1992), Davidson and MacKinnon (1993), Mariano (2001) and Stock and Watson (2003). Additional references include the econometric texts cited in Chapter 3 and the following:

Anderson, T.W. and H. Rubin (1950), "The Asymptotic Properties of Estimates of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 21: 570-582.

Baltagi, B.H. (1989), "A Hausman Specification Test in a Simultaneous Equations Model," *Econometric Theory*, Solution 88.3.5, 5: 453-467.

Basmann, R.L. (1957), "A Generalized Classical Method of Linear Estimation of Coefficients in a Structural Equation," *Econometrica*, 25: 77-83.

Basmann, R.L. (1960), "On Finite Sample Distributions of Generalized Classical Linear Identifiability Tests Statistics," *Journal of the American Statistical Association*, 55: 650-659.

Bekker, P.A. (1994), "Alternative Approximations to the Distribution of Instrumental Variable Estimators," *Econometrica* 62: 657-681.

Bekker, P.A. and T.J. Wansbeek (2001), "Identification in Parametric Models," Chapter 7 in Baltagi, B.H. (ed.) *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Bound, J., D.A. Jaeger and R.M. Baker (1995), "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Exogenous Explanatory Variable is Weak," *Journal of American Statistical Association* 90: 443-450.

Cornwell, C. and W.N. Trumbull (1994), "Estimating the Economic Model of Crime Using Panel Data," *Review of Economics and Statistics*, 76: 360-366.

Cragg, J.G. and S.G. Donald (1996), "Inferring the Rank of Matrix," *Journal of Econometrics*, 76: 223-250.

Deaton, A. (1997), *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy* (Johns Hopkins University Press: Baltimore).

Durbin, J. (1954), "Errors in Variables," *Review of the International Statistical Institute*, 22: 23-32.

Farebrother, R.W. (1985), "The Exact Bias of Wald's Estimator," *Econometric Theory*, Problem 85.3.1, 1: 419.

Farebrother, R.W. (1991), "Comparison of t-Ratios," *Econometric Theory*, Solution 90.1.4, 7: 145-146.

Fisher, F.M. (1966), *The Identification Problem in Econometrics* (McGraw-Hill: New York).

Gao, C. and K. Lahiri (2000), "Degeneration of Feasible GLS to 2SLS in a Limited Information Simultaneous Equations Model," *Econometric Theory*, Problem 00.2.1, 16: 287.

Gronau, R. (1973), "The Effects of Children on the Housewife's Value of Time," *Journal of Political Economy*, 81: S168-199.

Haavelmo, T. (1944), "The Probability Approach in Econometrics," *Supplement to Econometrica*, 12.

Hall, A. (1993), "Some Aspects of Generalized Method of Moments Estimation," Chapter 15 in *Handbook of Statistics*, Vol. 11 (North Holland: Amsterdam).

Hansen, L. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50: 646-660.

Hausman, J.A. (1978), "Specification Tests in Econometrics," *Econometrica*, 46: 1251-1272.

Hausman, J.A. (1983), "Specification and Estimation of Simultaneous Equation Models," Chapter 7 in Griliches, Z. and Intriligator, M.D. (eds.) *Handbook of Econometrics*, Vol. I (North Holland: Amsterdam).

Holly, A. (1987), "Identification and Estimation of a Simple Two-Equation Model," *Econometric Theory*, Problem 87.3.3, 3: 463-466.

Holly, A. (1988), "A Hausman Specification Test in a Simultaneous Equations Model," *Econometric Theory*, Problem 88.3.5, 4: 537-538.

Holly, A. (1990), "Comparison of t-ratios," *Econometric Theory*, Problem 90.1.4, 6: 114.

Kapteyn, A. and D.G. Fiebig (1981), "When are Two-Stage and Three-Stage Least Squares Estimators Identical?," *Economics Letters*, 8: 53-57.

Koopmans, T.C. and J. Marschak (1950), *Statistical Inference in Dynamic Economic Models* (John Wiley and Sons: New York).

Koopmans, T.C. and W.C. Hood (1953), *Studies in Econometric Method* (John Wiley and Sons: New York).

Laffer, A.B., (1970), "Trade Credit and the Money Market," *Journal of Political Economy*, 78: 239-267.

Lott, W.F. and S.C. Ray (1992), *Applied Econometrics: Problems with Data Sets* (The Dryden Press: New York).

Manski, C.F. (1995), *Identification Problems in the Social Sciences* (Harvard University Press: Cambridge)

Mariano, R.S. (2001), "Simultaneous Equation Model Estimators: Statistical Properties and Practical Implications," Chapter 6 in Baltagi, B.H. (ed.) *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Nelson, C.R. and R. Startz (1990), "The Distribution of the Instrumental Variables Estimator and its t-Ratio when the Instrument is a Poor One," *Journal of Business*, 63: S125-140.

Sapra, S.K. (1997), "Equivariance of an Instrumental Variable (IV) Estimator in the Linear Regression Model," *Econometric Theory*, Problem 97.2.5, 13: 464.

Singh, N. and A N. Bhat (1988), "Identification and Estimation of a Simple Two-Equation Model," *Econometric Theory*, Solution 87.3.3, 4: 542-545.

Staiger, D. and J. Stock (1997), "Instrumental Variables Regression With Weak Instruments," *Econometrica,* 65: 557-586.

Stock, J.H. and M.W. Watson (2003), *Introduction to Econometrics* (Addison Wesley: Boston).

Theil, H. (1953), "Repeated Least Squares Applied to Complete Equation Systems," *The Hague, Central Planning Bureau* (Mimeo).

Theil, H. (1971), *Principles of Econometrics* (Wiley: New York).

Wooldridge, J.M. (1990), "A Note on the Lagrange Multiplier and F Statistics for Two Stage Least Squares Regression," *Economics Letters*, 34: 151-155.

Wald, A. (1940), "Fitting of Straight Lines if Both Variables are Subject to Error," *Annals of Mathematical Statistics*, 11: 284-300.

Wu, D.M. (1973), "Alternative Tests of Independence Between Stochastic Regressors and Disturbances," *Econometrica*, 41: 733-740.

Zellner, A. and Theil, H. (1962), "Three-Stage Least Squares: Simultaneous Estimation of Simultaneous Equations," *Econometrica*, 30: 54-78.

Appendix: The Identification Problem Revisited: The Rank Condition of Identification

In section 11.1.2, we developed a *necessary* but not sufficient condition for identification. In this section we emphasize that model identification is crucial because only then can we get meaningful estimates of the parameters. For an *under-identified* model, different sets of parameter values agree well with the statistical evidence. As Bekker and Wansbeek (2001, p. 144) put it, preference for one set of parameter values over other ones becomes arbitrary. Therefore, "Scientific conclusions drawn on the basis of such arbitrariness are in the best case void and in the worst case dangerous." Manski (1995, p. 6) also warns that "negative identification findings imply that statistical inference is fruitless. It makes no sense to try to use a sample of finite size to infer something that could not be learned even if a sample of infinite size were available."

Consider the simultaneous equation model

$$By_t + \Gamma x_t = u_t \qquad t = 1, 2, \ldots, T. \tag{A.1}$$

which displays the whole set of equations at time t. B is $G \times G$, Γ is $G \times K$ and u_t is $G \times 1$. B is square and nonsingular indicating that the system is complete, i.e., there are as many equations as there are endogenous variables. Premultiplying (A.1) by B^{-1} and solving for y_t in terms of the exogenous variables and the vector of errors, we get

$$y_t = \Pi x_t + v_t \quad t = 1, 2, \ldots, T. \tag{A.2}$$

where $\Pi = -B^{-1}\Gamma$, is $G \times K$, and $v_t = B^{-1}u_t$. Note that if we premultiply the structural model in (A.1) by an arbitrary nonsingular $G \times G$ matrix F, then the new structural model has the same reduced form given in (A.2). In this case, each new structural equation is a linear combination of the original structural equations, but the reduced form equations are the same. One idea for identification, explored by Fisher (1966), is to note that (A.1) is completely defined by B, Γ and the probability density function of the disturbances $p(u_t)$. The specification of the structural model which comes from economic theory, imposes a lot of zero restrictions on the B and Γ coefficients. In addition, there may be cross-equations or within equation restrictions. For example, constant returns to scale of the production function, or homogeneity of a demand equation, or symmetry conditions. In addition, the probability density function of the disturbances may itself contain some zero covariance restrictions. The structural model given in (A.1) is identified if these restrictions are enough to distinguish it from any other structural model. This is operationalized by proving that the only nonsingular matrix F which results in a new structural model that satisfies the same restrictions on the original model is the identity matrix. If after imposing the restrictions, only certain rows of F resemble the corresponding rows of an identity matrix, up to a scalar of normalization, then the corresponding equations of the system are identified. The remaining equations are not identified. This is the same concept of taking a linear combination of the demand and supply equations and seeing if the linear combination is different from demand or supply. If it is, then both equations are identified. If it looks like demand but not supply, then supply is identified and demand is not identified. Let us look at an example.

Example (A.1): Consider a demand and supply equations with

$$Q_t = a - bP_t + cY_t + u_{1t} \tag{A.3}$$

$$Q_t = d + eP_t + fW_t + u_{2t} \tag{A.4}$$

where Y is income and W is weather. Writing (A.3) and (A.4) in matrix form (A.1), we get

$$B = \begin{bmatrix} 1 & b \\ 1 & -e \end{bmatrix} \qquad \Gamma = \begin{bmatrix} -a & -c & 0 \\ -d & 0 & -f \end{bmatrix} \qquad y_t = \begin{pmatrix} Q_t \\ P_t \end{pmatrix} \tag{A.5}$$

$$x_t' = [1, Y_t, W_t] \qquad u_t' = (u_{1t}, u_{2t})$$

There are two zero restrictions on Γ. The first is that income does not appear in the supply equation and the second is that weather does not appear in the demand equation. Therefore, the order condition of identification is satisfied for both equations. In fact, for each equation, there is one excluded exogenous variable and only one right hand side included endogenous variable. Therefore, both equations are just-identified. Let $F = [f_{ij}]$ for $i, j = 1, 2$, be a nonsingular matrix. Premultiply this system by F. The new matrix B is now FB and the new matrix Γ is now $F\Gamma$. In order for the transformed system to satisfy the same restrictions as the original model, FB must satisfy the following normalization restrictions:

$$f_{11} + f_{12} = 1 \qquad f_{21} + f_{22} = 1 \tag{A.6}$$

Also, $F\Gamma$ should satisfy the following zero restrictions

$$-f_{21}c + f_{22}0 = 0 \qquad f_{11}0 - f_{12}f = 0 \tag{A.7}$$

Since $c \neq 0$, and $f \neq 0$, then (A.7) implies that $f_{21} = f_{12} = 0$. Using (A.6), we get $f_{11} = f_{22} = 1$. Hence, the only nonsingular F that satisfies the same restrictions on the original model is the identity matrix, provided $c \neq 0$ and $f \neq 0$. Therefore, both equations are identified.

Example (A.2): If income does not appear in the demand equation (A.3), i.e., $c = 0$, the model looks like

$$Q_t = a - bP_t + u_{1t} \tag{A.8}$$
$$Q_t = d + eP_t + fW_t + u_{2t} \tag{A.9}$$

In this case, only the second restriction given in (A.7) holds. Therefore, $f \neq 0$ implies $f_{12} = 0$, however f_{21} is not necessarily zero without additional restrictions. Using (A.6), we get $f_{11} = 1$ and $f_{21} + f_{22} = 1$. This means that only the first row of F looks like the first row of an identity matrix, and only the demand equation is identified. In fact, the order condition for identification is not met for the supply equation since there are no excluded exogenous variables from that equation but there is one right hand side included endogenous variable. See problems 6 and 7 for more examples of this method of identification.

Example (A.3): Suppose that $u \sim (0, \Omega)$ where $\Omega = \Sigma \otimes I_T$, and $\Sigma = [\sigma_{ij}]$ for $i, j = 1, 2$. This example shows how a variance-covariance restriction can help identify an equation. Let us take the model defined in (A.8), (A.9) and add the restriction that $\sigma_{12} = \sigma_{21} = 0$. In this case, the transformed model disturbances will be $Fu \sim (0, \Omega^*)$, where $\Omega^* = \Sigma^* \otimes I_T$, and $\Sigma^* = F\Sigma F'$. In fact, since Σ is diagonal, $F\Sigma F'$ should also be diagonal. This imposes the following restriction on the elements of F:

$$f_{11}\sigma_{11}f_{21} + f_{12}\sigma_{22}f_{22} = 0 \tag{A.10}$$

But, $f_{11} = 1$ and $f_{12} = 0$ from the zero restrictions imposed on the demand equation, see example 4. Hence, (A.10) reduces to $\sigma_{11}f_{21} = 0$. Since $\sigma_{11} \neq 0$, this implies that $f_{21} = 0$, and the normalization restriction given in (A.6), implies that $f_{22} = 1$. Therefore, the second equation is also identified.

Example (A.4): In this example, we demonstrate how cross-equation restrictions can help identify equations. Consider the following simultaneous model

$$y_1 = a + by_2 + cx_1 + u_1 \tag{A.11}$$
$$y_2 = d + ey_1 + fx_1 + gx_2 + u_2 \tag{A.12}$$

and add the restriction $c = f$. It can be easily shown that the first equation is identified with $f_{11} = 1$, and $f_{12} = 0$. The second equation has no zero restrictions, but the cross-equation restriction $c = f$ implies:

$$-cf_{11} - ff_{12} = -cf_{21} - ff_{22}$$

Using $c = f$, we get

$$f_{11} + f_{12} = f_{21} + f_{22} \tag{A.13}$$

But, the first equation is identified with $f_{11} = 1$ and $f_{12} = 0$. Hence, (A.13) reduces to $f_{21} + f_{22} = 1$, which together with the normalization condition $-f_{21}b + f_{22} = 1$, gives $f_{21}(b + 1) = 0$. If $b \neq -1$, then $f_{21} = 0$ and $f_{22} = 1$. The second equation is also identified provided $b \neq -1$.

Alternatively, one can look at the problem of identification by asking whether the structural parameters in B and Γ can be obtained from the reduced form parameters. It will be clear from the following discussion that this task is impossible if there are no restrictions on this

simultaneous model. In this case, the model is hopelessly unidentified. However, in the usual case where there are a lot of zeros in B and Γ, we may be able to retrieve the remaining non-zero coefficients from Π. More rigorously, $\Pi = -B^{-1}\Gamma$, which implies that

$$B\Pi + \Gamma = 0 \tag{A.14}$$

or

$$AW = 0 \quad \text{where} \quad A = [B, \Gamma] \quad \text{and} \quad W' = [\Pi', I_K] \tag{A.15}$$

For the first equation, this implies that

$$\alpha_1' W = 0 \quad \text{where} \quad \alpha_1' \text{ is the first row of } A. \tag{A.16}$$

W is known (or can be estimated) and is of rank K. If the first equation has no restrictions on its structural parameters, then α_1' contains $(G+K)$ unknown coefficients. These coefficients satisfy K homogeneous equations given in (A.16). Without further restrictions, we cannot solve for $(G+K)$ coefficients (α_1') with only K equations. Let ϕ denote the matrix of R zero restrictions on the first equation, i.e., $\alpha_1'\phi = 0$. This together with (A.16) implies that

$$\alpha_1'[W, \phi] = 0 \tag{A.17}$$

and we can solve uniquely for α_1' (up to a scalar of normalization) provided the

$$\text{rank } [W, \phi] = G + K - 1 \tag{A.18}$$

Economists specify each structural equation with the left hand side endogenous variable having the coefficient one. This normalization identifies one coefficient of α_1', therefore, we require only $(G + K - 1)$ more restrictions to uniquely identify the remaining coefficients of α_1. $[W, \phi]$ is a $(G + K) \times (K + R)$ matrix. Its rank is less than any of its two dimensions, i.e., $(K + R) \geq (G + K - 1)$, which results in $R \geq G - 1$, or the *order condition* of identification. Note that this is a necessary but not sufficient condition for (A.18) to hold. It states that the number of restrictions on the first equation must be greater than the number of endogenous variables minus one. If all R restrictions are zero restrictions, then it means that the number of excluded exogenous plus the number of excluded endogenous variables should be greater than $(G - 1)$. But the G endogenous variables are made up of the left hand side endogenous variable y_1, the g_1 right hand side included endogenous variables Y_1, and $(G - g_1 - 1)$ excluded endogenous variables. Therefore, $R \geq (G - 1)$ can be written as $k_2 + (G - g_1 - 1) \geq (G - 1)$ which reduces to $k_2 \geq g_1$, which was discussed earlier in this chapter.

The *necessary and sufficient condition* for identification can now be obtained as follows: Using (A.1) one can write

$$Az_t = u_t \quad \text{where} \quad z_t' = (y_t', x_t') \tag{A.19}$$

and from the first definition of identification we make the transformation $FAz_t = Fu_t$, where F is a $G \times G$ nonsingular matrix. The first equation satisfies the restrictions $\alpha_1'\phi = 0$, which can be rewritten as $\iota' A\phi = 0$, where ι' is the first row of an identity matrix I_G. F must satisfy the restriction that (first row of $FA)\phi = 0$. But the first row of FA is the first row of F, say f_1', times A. This means that $f_1'(A\phi) = 0$. For the first equation to be identified, this condition

on the transformed first equation must be equivalent to $\iota' A\phi = 0$, up to a scalar constant. This holds if and only if f_1' is a scalar multiple of ι', and the latter condition holds if and only if the rank $(A\phi) = G - 1$. The latter is known as the *rank condition* for identification.

Example (A.5): Consider the simple Keynesian model given in example 1. The second equation is an identity and the first equation satisfies the order condition of identification, since I_t is the excluded exogenous variable from that equation, and there is only one right hand side included endogenous variable Y_t. In fact, the first equation is just-identified. Note that

$$A = [B, \Gamma] = \begin{bmatrix} \beta_{11} & \beta_{12} & \gamma_{11} & \gamma_{12} \\ \beta_{21} & \beta_{22} & \gamma_{21} & \gamma_{22} \end{bmatrix} = \begin{bmatrix} 1 & -\beta & -\alpha & 0 \\ -1 & 1 & 0 & -1 \end{bmatrix} \tag{A.20}$$

and ϕ for the first equation consists of only one restriction, namely that I_t is not in that equation, or $\gamma_{12} = 0$. This makes $\phi' = (0, 0, 0, 1)$, since $\alpha_1'\phi = 0$ gives $\gamma_{12} = 0$. From (A.20), $A\phi = (\gamma_{12}, \gamma_{22})' = (0, -1)'$ and the rank $(A\phi) = 1 = G - 1$. Hence, the rank condition holds for the first equation and it is identified. Problem 8 reconsiders example (A.1), where both equations are just-identified by the order condition of identification and asks the reader to show that both satisfy the rank condition of identification as long as $c \neq 0$ and $f \neq 0$.

The reduced form of the simple Keynesian model is given in equations (11.4) and (11.5). In fact,

$$\Pi = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} = \begin{bmatrix} \alpha & \beta \\ \alpha & 1 \end{bmatrix} / (1 - \beta \tag{A.21}$$

Note that

$$\begin{aligned} \pi_{11}/\pi_{22} &= \alpha \quad \text{and} \quad \pi_{21}/\pi_{22} = \alpha \\ \pi_{12}/\pi_{22} &= \beta \quad \text{and} \quad (\pi_{22} - 1)/\pi_{22} = \beta \end{aligned} \tag{A.22}$$

Therefore, the structural parameters of the consumption equation can be retrieved from the reduced form coefficients. However, what happens if we replace Π by its OLS estimate $\widehat{\Pi}_{OLS}$? Would the solution in (A.22) lead to two estimates of (α, β) or would this solution lead to a unique estimate? In this case, the consumption equation is just-identified and the solution in (A.22) is unique. To show this, recall that

$$\widehat{\pi}_{12} = m_{ci}/m_{ii}; \quad \text{and} \quad \widehat{\pi}_{22} = m_{yi}/m_{ii} \tag{A.23}$$

Solving for $\widehat{\beta}$, using (A.22), one gets

$$\widehat{\beta} = \widehat{\pi}_{12}/\widehat{\pi}_{22} = m_{ci}/m_{yi} \tag{A.24}$$

and

$$\widehat{\beta} = (\widehat{\pi}_{22} - 1)/\widehat{\pi}_{22} = (m_{ci} - m_{yi})/m_{yi} \tag{A.25}$$

(A.24) and (A.25) lead to a unique solution because equation (11.2) gives

$$m_{yi} = m_{ci} + m_{ii} \tag{A.26}$$

In general, we would not be able to solve for the structural parameters of an unidentified equation in terms of the reduced form parameters. However, when this equation is identified,

replacing the reduced form parameters by their OLS estimates would lead to a unique estimate of the structural parameters, only if this equation is just-identified, and to more than one estimate depending upon the degree of over-identification. Problem 8 gives another example of the just-identified case, while problem 9 considers a model with one unidentified and another over-identified equation.

Example (A.6): Equations (11.13) and (11.14) give an unidentified demand and supply model with

$$B = \begin{bmatrix} 1 & -\beta \\ 1 & -\delta \end{bmatrix} \quad \text{and} \quad \Gamma = \begin{bmatrix} -\alpha \\ -\gamma \end{bmatrix} \tag{A.27}$$

The reduced form equations given by (11.16) and (11.17) yield

$$\Pi = -B^{-1}\Gamma = \begin{bmatrix} \pi_{11} \\ \pi_{21} \end{bmatrix} = \begin{bmatrix} \alpha\delta - \gamma\beta \\ \alpha - \gamma \end{bmatrix} / (\delta - \beta) \tag{A.28}$$

Note that one cannot solve for (α, β) nor (γ, δ) in terms of (π_{11}, π_{21}) without further restrictions.

CHAPTER 12
Pooling Time-Series of Cross-Section Data

12.1 Introduction

In this chapter, we will consider pooling time-series of cross-sections. This may be a panel of households or firms or simply countries or states followed over time. Two well known examples of panel data in the U.S. are the Panel Study of Income Dynamics (PSID) and the National Longitudinal Survey (NLS). The PSID began in 1968 with 4802 families, including an over-sampling of poor households. Annual interviews were conducted and socioeconomic characteristics of each of the families and of roughly 31000 individuals who have been in these or derivative families were recorded. The list of variables collected is over 5000. The NLS, followed five distinct segments of the labor force. The original samples include 5020 older men, 5225 young men, 5083 mature women, 5159 young women and 12686 youths. There was an over-sampling of blacks, hispanics, poor whites and military in the youths survey. The list of variables collected runs into the thousands. An inventory of national studies using panel data is given at http://www.isr.umich.edu/src/psid/panelstudies.html. Pooling this data gives a richer source of variation which allows for more efficient estimation of the parameters. With additional, more informative data, one can get more reliable estimates and test more sophisticated behavioral models with less restrictive assumptions. Another advantage of panel data sets are their ability to control for individual heterogeneity. Not controlling for these unobserved individual specific effects leads to bias in the resulting estimates. Panel data sets are also better able to identify and estimate effects that are simply not detectable in pure cross-sections or pure time-series data. In particular, panel data sets are better able to study complex issues of dynamic behavior. For example, with a cross-section data set one can estimate the rate of unemployment at a particular point in time. Repeated cross-sections can show how this proportion changes over time. Only panel data sets can estimate what proportion of those who are unemployed in one period remain unemployed in another period. Some of the benefits and limitations of using panel data sets are listed in Hsiao (2003) and Baltagi (2005). Section 12.2 studies the error components model focusing on fixed effects, random effects and maximum likelihood estimation. Section 12.3 considers the question of prediction in a random effects model, while Section 12.4 illustrates the estimation methods using an empirical example. Section 12.5 considers testing the poolability assumption, the existence of random individual effects and the consistency of the random effects estimator using a Hausman test. Section 12.6 studies the dynamic panel data model and illustrates the methods used with an empirical example. Section 12.7 concludes with a short presentation of program evaluation and the difference-in-differences estimator.

12.2 The Error Components Model

The regression model is still the same, but it now has double subscripts

$$y_{it} = \alpha + X'_{it}\beta + u_{it} \tag{12.1}$$

where i denotes cross-sections and t denotes time-periods with $i = 1, 2, \ldots, N$, and $t = 1, 2, \ldots, T$. α is a scalar, β is $K \times 1$ and X_{it} is the it-th observation on K explanatory variables. The observations are usually stacked with i being the slower index, i.e., the T observations on the first household followed by the T observations on the second household, and so on, until we get to the N-th household. Under the error components specification, the disturbances take the form

$$u_{it} = \mu_i + \nu_{it} \tag{12.2}$$

where the μ_i's are cross-section specific components and ν_{it} are remainder effects. For example, μ_i may denote individual ability in an earnings equation, or managerial skill in a production function or simply a country specific effect. These effects are time-invariant.

In vector form, (12.1) can be written as

$$y = \alpha \iota_{NT} + X\beta + u = Z\delta + u \tag{12.3}$$

where y is $NT \times 1$, X is $NT \times K$, $Z = [\iota_{NT}, X]$, $\delta' = (\alpha', \beta')$, and ι_{NT} is a vector of ones of dimension NT. Also, (12.2) can be written as

$$u = Z_\mu \mu + \nu \tag{12.4}$$

where $u' = (u_{11}, \ldots, u_{1T}, u_{21}, \ldots, u_{2T}, \ldots, u_{N1}, \ldots, u_{NT})$ and $Z_\mu = I_N \otimes \iota_T$. I_N is an identity matrix of dimension N, ι_T is a vector of ones of dimension T, and $\otimes$ denotes Kronecker product defined in the Appendix to Chapter 7. Z_μ is a selector matrix of ones and zeros, or simply the matrix of individual dummies that one may include in the regression to estimate the μ_i's if they are assumed to be fixed parameters. $\mu' = (\mu_1, \ldots, \mu_N)$ and $\nu' = (\nu_{11}, \ldots, \nu_{1T}, \ldots, \nu_{N1}, \ldots, \nu_{NT})$. Note that $Z_\mu Z_\mu' = I_N \otimes J_T$ where J_T is a matrix of ones of dimension T, and $P = Z_\mu (Z_\mu' Z_\mu)^{-1} Z_\mu'$, the projection matrix on Z_μ, reduces to $P = I_N \otimes \bar{J}_T$ where $\bar{J}_T = J_T/T$. P is a matrix which averages the observation across time for each individual, and $Q = I_{NT} - P$ is a matrix which obtains the deviations from individual means. For example, Pu has a typical element $\bar{u}_{i.} = \sum_{t=1}^{T} u_{it}/T$ repeated T times for each individual and Qu has a typical element $(u_{it} - \bar{u}_{i.})$. P and Q are (i) symmetric *idempotent* matrices, i.e., $P' = P$ and $P^2 = P$. This means that the rank $(P) = \text{tr}(P) = N$ and rank $(Q) = \text{tr}(Q) = N(T-1)$. This uses the result that rank of an idempotent matrix is equal to its trace, see Graybill (1961, Theorem 1.63) and the Appendix to Chapter 7. Also, (ii) P and Q are *orthogonal*, i.e., $PQ = 0$ and (iii) they *sum to the identity matrix* $P + Q = I_{NT}$. In fact, any two of these properties imply the third, see Graybill (1961, Theorem 1.68).

12.2.1 The Fixed Effects Model

If the μ_i's are thought of as *fixed* parameters to be estimated, then equation (12.1) becomes

$$y_{it} = \alpha + X_{it}'\beta + \sum_{i=1}^{N} \mu_i D_i + \nu_{it} \tag{12.5}$$

where D_i is a dummy variable for the i-th household. Not all the dummies are included so as not to fall in the dummy variable trap. One is usually dropped or equivalently, we can say that there is a restriction on the μ's given by $\sum_{i=1}^{N} \mu_i = 0$. The ν_{it}'s are the usual classical IID random variables with 0 mean and variance σ_ν^2. OLS on equation (12.5) is BLUE, but we have two

problems, the first is the loss of degrees of freedom since in this case, we are estimating $N+K$ parameters. Also, with a lot of dummies we could be running into multicollinearity problems and a large $X'X$ matrix to invert. For example, if $N=50$ states, $T=10$ years and we have two explanatory variables, then with 500 observations we are estimating 52 parameters. Alternatively, we can think of this in an analysis of variance context and rearrange our observations, say, on y in an $(N \times T)$ matrix where rows denote firms and columns denote time periods.

		t				
		1	2	..	T	
	1	y_{11}	y_{12}	..	y_{1T}	$y_{1.}$
i	2	y_{21}	y_{22}	..	y_{2T}	$y_{2.}$
	:	:	:	..	:	:
	N	y_{N1}	y_{N2}	..	y_{NT}	$y_{N.}$

where $y_{i.} = \sum_{t=1}^{T} y_{it}$ and $\bar{y}_{i.} = y_{i.}/T$. For the simple regression with one regressor, the model given in (12.1) becomes

$$y_{it} = \alpha + \beta x_{it} + \mu_i + \nu_{it} \tag{12.6}$$

averaging over time gives

$$\bar{y}_{i.} = \alpha + \beta \bar{x}_{i.} + \mu_i + \bar{\nu}_{i.} \tag{12.7}$$

and averaging over all observations gives

$$\bar{y}_{..} = \alpha + \beta \bar{x}_{..} + \bar{\nu}_{..} \tag{12.8}$$

where $\bar{y}_{..} = \sum_{i=1}^{N} \sum_{t=1}^{T} y_{it}/NT$. Equation (12.8) follows because the μ_i's sum to zero. Defining $\widetilde{y}_{it} = (y_{it} - \bar{y}_{i.})$ and $\widetilde{x}_{it}$ and $\widetilde{\nu}_{it}$ similarly, we get

$$y_{it} - \bar{y}_{i.} = \beta(x_{it} - \bar{x}_{i.}) + (\nu_{it} - \bar{\nu}_{i.})$$

or

$$\widetilde{y}_{it} = \beta \widetilde{x}_{it} + \widetilde{\nu}_{it} \tag{12.9}$$

Running OLS on equation (12.9) leads to the same estimator of β as that obtained from equation (12.5). This is called the least squares dummy variable estimator (LSDV) or $\widetilde{\beta}$ in our notation. It is also known as the *Within estimator* since $\sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{x}_{it}^2$ is the *within* sum of squares in an analysis of variance framework. One can then retrieve an estimate of α from equation (12.8) as $\widetilde{\alpha} = \bar{y}_{..} - \widetilde{\beta} \bar{x}_{..}$. Similarly, if we are interested in the μ_i's, those can also be retrieved from (12.7) and (12.8) as follows:

$$\widetilde{\mu}_i = (\bar{y}_{i.} - \bar{y}_{..}) - \widetilde{\beta}(\bar{x}_{i.} - \bar{x}_{..}) \tag{12.10}$$

In matrix form, one can substitute the disturbances given by (12.4) into (12.3) to get

$$y = \alpha \iota_{NT} + X\beta + Z_\mu \mu + \nu = Z\delta + Z_\mu \mu + \nu \tag{12.11}$$

and then perform OLS on (12.11) to get estimates of α, β and μ. Note that Z is $NT \times (K+1)$ and Z_μ, the matrix of individual dummies is $NT \times N$. If N is large, (12.11) will include too

many individual dummies, and the matrix to be inverted by OLS is large and of dimension $(N+K)$. In fact, since α and β are the parameters of interest, one can obtain the least squares dummy variables (LSDV) estimator from (12.11), by residualing out the Z_μ variables, i.e., by premultiplying the model by Q, the orthogonal projection of Z_μ, and performing OLS

$$Qy = QX\beta + Q\nu \tag{12.12}$$

This uses the fact that $QZ_\mu = Q\iota_{NT} = 0$, since $PZ_\mu = Z_\mu$. In other words, the Q matrix wipes out the individual effects. Recall, the FWL Theorem in Chapter 7. This is a regression of $\widetilde{y} = Qy$ with typical element $(y_{it} - \bar{y}_{i.})$ on $\widetilde{X} = QX$ with typical element $(X_{it,k} - \bar{X}_{i.,k})$ for the k-th regressor, $k = 1, 2, \ldots, K$. This involves the inversion of a $(K \times K)$ matrix rather than $(N+K) \times (N+K)$ as in (12.11). The resulting OLS estimator is

$$\widetilde{\beta} = (X'QX)^{-1}X'Qy \tag{12.13}$$

with $\text{var}(\widetilde{\beta}) = \sigma^2_\nu (X'QX)^{-1} = \sigma^2_\nu (\widetilde{X}'\widetilde{X})^{-1}$.

Note that this fixed effects (FE) estimator cannot estimate the effect of any time-invariant variable like sex, race, religion, schooling, or union participation. These time-invariant variables are wiped out by the Q transformation, the deviations from means transformation. Alternatively, one can see that these time-invariant variables are spanned by the individual dummies in (12.5) and therefore any regression package attempting (12.5) will fail, signaling perfect multicollinearity. If (12.5) is the true model, LSDV is BLUE as long as ν_{it} is the standard classical disturbance with mean 0 and variance covariance matrix $\sigma^2_\nu I_{NT}$. Note that as $T \to \infty$, the FE estimator is consistent. However, if T is fixed and $N \to \infty$ as typical in short labor panels, then only the FE estimator of β is consistent, the FE estimators of the individual effects $(\alpha + \mu_i)$ are not consistent since the number of these parameters increase as N increases.

Testing for Fixed Effects: One could test the joint significance of these dummies, i.e., H_0; $\mu_1 = \mu_2 = .. = \mu_{N-1} = 0$, by performing an F-test. This is a simple Chow test given in (4.17) with the restricted residual sums of squares (RRSS) being that of OLS on the pooled model and the unrestricted residual sums of squares (URSS) being that of the LSDV regression. If N is large, one can perform the within transformation and use that residual sum of squares as the URSS. In this case

$$F_0 = \frac{(RRSS - URSS)/(N-1)}{URSS/(NT-N-K)} \overset{H_0}{\sim} F_{N-1,N(T-1)-K} \tag{12.14}$$

Computational Warning: One computational caution for those using the *Within* regression given by (12.12). The s^2 of this regression as obtained from a typical regression package divides the residual sums of squares by $NT - K$ since the intercept and the dummies are not included. The proper s^2, say s^{*2} from the LSDV regression in (12.5) would divide the same residual sums of squares by $N(T-1) - K$. Therefore, one has to adjust the variances obtained from the within regression (12.12) by multiplying the variance-covariance matrix by (s^{*2}/s^2) or simply by multiplying by $[NT - K]/[N(T-1) - K]$.

12.2.2 The Random Effects Model

There are too many parameters in the fixed effects model and the loss of degrees of freedom can be avoided if the μ_i's can be assumed random. In this case $\mu_i \sim \text{IID}(0, \sigma^2_\mu)$, $\nu_{it} \sim \text{IID}(0, \sigma^2_\nu)$

and the μ_i's are independent of the ν_{it}'s. In addition, the X_{it}'s are independent of the μ_i's and ν_{it}'s for all i and t. The random effects model is an appropriate specification if we are drawing N individuals randomly from a large population.

This specification implies a homoskedastic variance $\text{var}(u_{it}) = \sigma_\mu^2 + \sigma_\nu^2$ for all i and t, and an equi-correlated block-diagonal covariance matrix which exhibits serial correlation over time only between the disturbances of the same individual. In fact,

$$\begin{aligned} \text{cov}(u_{it}, u_{js}) &= \sigma_\mu^2 + \sigma_\nu^2 \quad \text{for} \quad i = j,\ t = s \\ &= \sigma_\mu^2 \qquad\quad \text{for} \quad i = j,\ t \neq s \end{aligned} \tag{12.15}$$

and zero otherwise. This also means that the correlation coefficient between u_{it} and u_{js} is

$$\begin{aligned} \rho &= \text{correl}(u_{it}, u_{js}) = 1 \quad \text{for} \quad i = j,\ t = s \\ &= \sigma_\mu^2/(\sigma_\mu^2 + \sigma_\nu^2) \quad \text{for} \quad i = j,\ t \neq s \end{aligned} \tag{12.16}$$

and zero otherwise. From (12.4), one can compute the variance-covariance matrix

$$\Omega = E(uu') = Z_\mu E(\mu\mu')Z_\mu' + E(\nu\nu') = \sigma_\mu^2(I_N \otimes J_T) + \sigma_\nu^2(I_N \otimes I_T) \tag{12.17}$$

In order to obtain the GLS estimator of the regression coefficients, we need Ω^{-1}. This is a huge matrix for typical panels and is of dimension $(NT \times NT)$. No brute force inversion should be attempted even if the researcher's application has a small N and T. For example, if we observe $N = 20$ firms over $T = 5$ time periods, Ω will be 100 by 100. We will follow a simple trick devised by Wansbeek and Kapteyn (1982) that allows the deviation of Ω^{-1} and $\Omega^{-1/2}$. Essentially, one replaces J_T by $T\bar{J}_T$, and I_T by $(E_T + \bar{J}_T)$ where E_T is by definition $(I_T - \bar{J}_T)$. In this case:

$$\Omega = T\sigma_\mu^2(I_N \otimes \bar{J}_T) + \sigma_\nu^2(I_N \otimes E_T) + \sigma_\nu^2(I_N \otimes \bar{J}_T)$$

collecting terms with the same matrices, we get

$$\Omega = (T\sigma_\mu^2 + \sigma_\nu^2)(I_N \otimes \bar{J}_T) + \sigma_\nu^2(I_N \otimes E_T) = \sigma_1^2 P + \sigma_\nu^2 Q \tag{12.18}$$

where $\sigma_1^2 = T\sigma_\mu^2 + \sigma_\nu^2$. (12.18) is the spectral decomposition representation of Ω, with σ_1^2 being the first unique characteristic root of Ω of multiplicity N and σ_ν^2 is the second unique characteristic root of Ω of multiplicity $N(T-1)$. It is easy to verify, using the properties of P and Q, that

$$\Omega^{-1} = \frac{1}{\sigma_1^2} P + \frac{1}{\sigma_\nu^2} Q \tag{12.19}$$

and

$$\Omega^{-1/2} = \frac{1}{\sigma_1} P + \frac{1}{\sigma_\nu} Q \tag{12.20}$$

In fact, $\Omega^r = (\sigma_1^2)^r P + (\sigma_\nu^2)^r Q$ where r is an arbitrary scalar. Now we can obtain GLS as a weighted least squares. Fuller and Battese (1974) suggested premultiplying the regression equation given in (12.3) by $\sigma_\nu \Omega^{-1/2} = Q + (\sigma_\nu/\sigma_1)P$ and performing OLS on the resulting transformed regression. In this case, $y^* = \sigma_\nu \Omega^{-1/2} y$ has a typical element $y_{it} - \theta \bar{y}_{i.}$ where $\theta = 1 - (\sigma_\nu/\sigma_1)$. This transformed regression inverts a matrix of dimension $(K+1)$ and can be easily implemented using any regression package.

The Best Quadratic Unbiased (BQU) estimators of the variance components arise naturally from the spectral decomposition of Ω. In fact, $Pu \sim (0, \sigma_1^2 P)$ and $Qu \sim (0, \sigma_\nu^2 Q)$ and

$$\widehat{\sigma}_1^2 = \frac{u'Pu}{\text{tr}(P)} = T \sum_{i=1}^{N} \bar{u}_{i.}^2 / N \tag{12.21}$$

and

$$\widehat{\sigma}_\nu^2 = \frac{u'Qu}{\text{tr}(Q)} = T \sum_{i=1}^{N} \sum_{t=1}^{T} (u_{it} - \bar{u}_{i.})^2 / N(T-1) \tag{12.22}$$

provide the BQU estimators of σ_1^2 and σ_ν^2, respectively, see Balestra (1973).

These are analysis of variance type estimators of the variance components and are MVU under normality of the disturbances, see Graybill (1961). The true disturbances are not known and therefore (12.21) and (12.22) are not feasible. Wallace and Hussain (1969) suggest substituting OLS residuals $\widehat{u}_{OLS}$ instead of the true u's. After all, the OLS estimates are still unbiased and consistent, but no longer efficient. Amemiya (1971) shows that these estimators of the variance components have a different asymptotic distribution from that knowing the true disturbances. He suggests using the LSDV residuals instead of the OLS residuals. In this case $\widetilde{u} = y - \widetilde{\alpha}\iota_{NT} - X\widetilde{\beta}$ where $\widetilde{\alpha} = \bar{y}_{..} - \bar{X}'_{..}\widetilde{\beta}$ and $\bar{X}'_{..}$ is a $1 \times K$ vector of averages of all regressors. Substituting these $\widetilde{u}$'s for u in (12.21) and (12.22) we get the Amemiya-type estimators of the variance components. The resulting estimates of the variance components have the same asymptotic distribution as that knowing the true disturbances.

Swamy and Arora (1972) suggest running two regressions to get estimates of the variance components from the corresponding mean square errors of these regressions. The first regression is the *Within* regression, given in (12.12), which yields the following s^2:

$$\widehat{\widehat{\sigma}}_\nu^2 = [y'Qy - y'QX(X'QX)^{-1}X'Qy]/[N(T-1) - K] \tag{12.23}$$

The second regression is the *Between* regression which runs the regression of averages across time, i.e.,

$$\bar{y}_{i.} = \alpha + \bar{X}'_{i.}\beta + \bar{u}_{i.} \qquad i = 1, \ldots, N \tag{12.24}$$

This is equivalent to premultiplying the model in (12.11) by P and running OLS. The only caution is that the latter regression has NT observations because it repeats the averages T times for each individual, while the cross-section regression in (12.24) is based on N observations. To remedy this, one can run the cross-section regression

$$y_{i.}/\sqrt{T} = \alpha(\sqrt{T}) + (X'_{i.}/\sqrt{T})\beta + u_{i.}/\sqrt{T} \tag{12.25}$$

where one can easily verify that $\text{var}(u_{i.}/\sqrt{T}) = \sigma_1^2$. This regression will yield an s^2 given by

$$\widehat{\widehat{\sigma}}_1^2 = (y'Py - y'PZ(Z'PZ)^{-1}Z'Py)/(N - K - 1) \tag{12.26}$$

Note that stacking the following two transformed regressions we just performed yields

$$\begin{pmatrix} Qy \\ Py \end{pmatrix} = \begin{pmatrix} QZ \\ PZ \end{pmatrix} \delta + \begin{pmatrix} Qu \\ Pu \end{pmatrix} \tag{12.27}$$

and the transformed error has mean 0 and variance-covariance matrix given by

$$\begin{pmatrix} \sigma_\nu^2 Q & 0 \\ 0 & \sigma_1^2 P \end{pmatrix}$$

Problem 6 asks the reader to verify that OLS on this system of $2NT$ observations yields OLS on the pooled model (12.3). Also, GLS on this system yields GLS on (12.3). Alternatively, one could get rid of the constant α by running the following stacked regressions:

$$\begin{pmatrix} Qy \\ (P-\bar{J}_{NT})y \end{pmatrix} = \begin{pmatrix} QX \\ (P-\bar{J}_{NT})X \end{pmatrix} \beta + \begin{pmatrix} Qu \\ (P-\bar{J}_{NT})u \end{pmatrix} \tag{12.28}$$

This follows from the fact the $Q\iota_{NT} = 0$ and $(P-\bar{J}_{NT})\iota_{NT} = 0$. The transformed error has zero mean and variance-covariance matrix

$$\begin{pmatrix} \sigma_\nu^2 Q & 0 \\ 0 & \sigma_1^2 (P-\bar{J}_{NT}) \end{pmatrix} \tag{12.29}$$

OLS on this system, yields OLS on (12.3) and GLS on (12.28) yields GLS on (12.3). In fact,

$$\begin{aligned} \widehat{\beta}_{GLS} &= [(X'QX/\sigma_\nu^2) + X'(P-\bar{J}_{NT})X/\sigma_1^2]^{-1}[(X'Qy/\sigma_\nu^2) + (X'(P-\bar{J}_{NT})y/\sigma_1^2)] \\ &= [W_{XX} + \phi^2 B_{XX}]^{-1}[W_{Xy} + \phi^2 B_{Xy}] \end{aligned} \tag{12.30}$$

with $\text{var}(\widehat{\beta}_{GLS}) = \sigma_\nu^2[W_{XX} + \phi^2 B_{XX}]^{-1}$. Note that $W_{XX} = X'QX$, $B_{XX} = X'(P-\bar{J}_{NT})X$ and $\phi^2 = \sigma_\nu^2/\sigma_1^2$. Also, the Within estimator of β is $\widetilde{\beta}_{Within} = W_{XX}^{-1} W_{Xy}$ and the Between estimator $\widehat{\beta}_{Between} = B_{XX}^{-1} B_{Xy}$. This shows that $\widehat{\beta}_{GLS}$ is a matrix weighted average of $\widetilde{\beta}_{Within}$ and $\widehat{\beta}_{Between}$ weighing each estimate by the inverse of its corresponding variance. In fact

$$\widehat{\beta}_{GLS} = W_1 \widetilde{\beta}_{Within} + W_2 \widehat{\beta}_{Between} \tag{12.31}$$

where $W_1 = [W_{XX} + \phi^2 B_{XX}]^{-1} W_{XX}$ and $W_2 = [W_{XX} + \phi^2 B_{XX}]^{-1}(\phi^2 B_{XX}) = I - W_1$. This was demonstrated by Maddala (1971). Note that (i) if $\sigma_\mu^2 = 0$, then $\phi^2 = 1$ and $\widehat{\beta}_{GLS}$ reduces to $\widehat{\beta}_{OLS}$. (ii) If $T \to \infty$, then $\phi^2 \to 0$ and $\widehat{\beta}_{GLS}$ tends to $\widetilde{\beta}_{Within}$. (iii) If $\phi^2 \to \infty$, then $\widehat{\beta}_{GLS}$ tends to $\widehat{\beta}_{Between}$. In other words, the Within estimator ignores the between variation, and the Between estimator ignores the within variation. The OLS estimator gives equal weight to the between and within variations. From (12.30), it is clear that $\text{var}(\widetilde{\beta}_{Within}) - \text{var}(\widehat{\beta}_{GLS})$ is a positive semi-definite matrix, since ϕ^2 is positive. However as $T \to \infty$ for any fixed N, $\phi^2 \to 0$ and both $\widehat{\beta}_{GLS}$ and $\widetilde{\beta}_{Within}$ have the same asymptotic variance.

Another estimator of the variance components was suggested by Nerlove (1971). His suggestion is to estimate $\widehat{\sigma}_\mu^2 = \sum_{i=1}^N (\widehat{\mu}_i - \overline{\widehat{\mu}})^2/(N-1)$ where $\widehat{\mu}_i$ are the dummy coefficients estimates from the LSDV regression. $\widehat{\sigma}_\nu^2$ is estimated from the within residual sums of squares divided by NT without correction for degrees of freedom.

Note that, except for Nerlove's (1971) method, one has to retrieve $\widehat{\sigma}_\mu^2$ as $(\widehat{\sigma}_1^2 - \widehat{\sigma}_\nu^2)/T$. In this case, there is no guarantee that the estimate of $\widehat{\sigma}_\mu^2$ would be non-negative. Searle (1971) has an extensive discussion of the problem of negative estimates of the variance components in the biometrics literature. One solution is to replace these negative estimates by zero. This in fact is the suggestion of the Monte Carlo study by Maddala and Mount (1973). This study finds that negative estimates occurred only when the true σ_μ^2 was small and close to zero. In these cases

OLS is still a viable estimator. Therefore, replacing negative $\widehat{\sigma}^2_\mu$ by zero is not a bad sin after all, and the problem is dismissed as not being serious.

Under the random effects model, GLS based on the true variance components is BLUE, and all the feasible GLS estimators considered are asymptotically efficient as either N or $T \to \infty$. Maddala and Mount (1973) compared OLS, Within, Between, feasible GLS methods, true GLS and MLE using their Monte Carlo study. They found little to choose among the various feasible GLS estimators in small samples and argued in favor of methods that were easier to compute.

Taylor (1980) derived exact finite sample results for the one-way error components model. He compared the Within estimator with the Swamy-Arora feasible GLS estimator. He found the following important results: (1) Feasible GLS is more efficient than FE for all but the fewest degrees of freedom. (2) The variance of feasible GLS is never more than 17% above the Cramér-Rao lower bound. (3) More efficient estimators of the variance components do not necessarily yield more efficient feasible GLS estimators. These finite sample results are confirmed by the Monte Carlo experiments carried out by Maddala and Mount (1973) and Baltagi (1981).

12.2.3 Maximum Likelihood Estimation

Under normality of the disturbances, one can write the log-likelihood function as

$$L(\alpha, \beta, \phi^2, \sigma^2_\nu) = \text{constant} - \frac{NT}{2}\log\sigma^2_\nu + \frac{N}{2}\log\phi^2 - \frac{1}{2\sigma^2_\nu}u'\Sigma^{-1}u \tag{12.32}$$

where $\Omega = \sigma^2_\nu\Sigma$, $\phi^2 = \sigma^2_\nu/\sigma^2_1$ and $\Sigma = Q + \phi^{-2}P$ from (12.18). This uses the fact that $|\Omega| =$ product of its characteristic roots $= (\sigma^2_\nu)^{N(T-1)}(\sigma^2_1)^N = (\sigma^2_\nu)^{NT}(\phi^2)^{-N}$. Note that there is a one-to-one correspondence between ϕ^2 and σ^2_μ. In fact, $0 \leq \sigma^2_\mu < \infty$ translates into $0 < \phi^2 \leq 1$. Brute force maximization of (12.32) leads to nonlinear first-order conditions, see Amemiya (1971). Instead, Breusch (1987) concentrates the likelihood with respect to α and σ^2_ν. In this case, $\widehat{\alpha}_{MLE} = \bar{y}_{..} - \bar{X}'_{..}\widehat{\beta}_{MLE}$ and $\widehat{\sigma}^2_{\nu,MLE} = \widehat{u}'\widehat{\Sigma}^{-1}\widehat{u}/NT$ where $\widehat{u}$ and $\widehat{\Sigma}$ are based on MLE's of β, ϕ^2 and α. Let $d = y - X\widehat{\beta}_{MLE}$ then $\widehat{\alpha}_{MLE} = \iota'_{NT}d/NT$ and $\widehat{u} = d - \iota_{NT}\widehat{\alpha} = d - \bar{J}_{NT}d$. This implies that $\widehat{\sigma}^2_{\nu,MLE}$ can be rewritten as

$$\widehat{\sigma}^2_{\nu,MLE} = d'[Q + \phi^2(P - \bar{J}_{NT})]d/NT \tag{12.33}$$

and the concentrated log-likelihood becomes

$$L_c(\beta, \phi^2) = \text{constant} - \frac{NT}{2}\log\{d'[Q + \phi^2(P - \bar{J}_{NT})]d\} + \frac{N}{2}\log\phi^2 \tag{12.34}$$

Maximizing (12.34), over ϕ^2 given β, yields

$$\widehat{\phi}^2 = \frac{d'Qd}{(T-1)d'(P - \bar{J}_{NT})d} = \frac{\sum_{i=1}^N\sum_{t=1}^T(d_{it} - \bar{d}_{i.})^2}{T(T-1)\sum_{i=1}^N(\bar{d}_{i.} - \bar{d}_{..})^2} \tag{12.35}$$

Maximizing (12.34) over β, given ϕ^2, yields

$$\widehat{\beta}_{MLE} = \left\{X'[Q + \phi^2(P - \bar{J}_{NT})]X\right\}^{-1}X'[Q + \phi^2(P - \bar{J}_{NT})]y \tag{12.36}$$

One can iterate between β and ϕ^2 until convergence. Breusch (1987) shows that provided $T > 1$, any i-th iteration β, call it β_i, gives $0 < \phi^2_{i+1} < \infty$ in the *(i+1)*th iteration. More importantly,

Breusch (1987) shows that these ϕ_i^2's have a *remarkable property* of forming a monotonic sequence. In fact, starting from the Within estimator of β, for $\phi^2 = 0$, the next ϕ^2 is finite and positive and starts a monotonically increasing sequence of ϕ^2's. Similarly, starting from the Between estimator of β, for $(\phi^2 \to \infty)$ the next ϕ^2 is finite and positive and starts a monotonically decreasing sequence of ϕ^2's. Hence, to guard against the possibility of a local maximum, Breusch (1987) suggests starting with $\widetilde{\beta}_{Within}$ and $\widehat{\beta}_{Between}$ and iterating. If these two sequences converge to the same maximum, then this is the global maximum. If one starts with $\widehat{\beta}_{OLS}$ for $\phi^2 = 1$, and the next iteration obtains a larger ϕ^2, then we have a local maximum at the boundary $\phi^2 = 1$. Maddala (1971) finds that there are at most two maxima for the likelihood $L(\phi^2)$ for $0 < \phi^2 \leq 1$. Hence, we have to guard against one local maximum.

12.3 Prediction

Suppose we want to predict S periods ahead for the i-th individual. For the random effects model, the BLU estimator is GLS. Using the results in Chapter 9 on GLS, Goldberger's (1962) Best Linear Unbiased Predictor (BLUP) of $y_{i,T+S}$ is

$$\hat{y}_{i,T+S} = Z'_{i,T+S}\widehat{\delta}_{GLS} + w'\Omega^{-1}\widehat{u}_{GLS} \quad \text{for } S \geq 1 \tag{12.37}$$

where $\widehat{u}_{GLS} = y - Z\widehat{\delta}_{GLS}$ and $w = E(u_{i,T+S}u)$. Note that

$$u_{i,T+S} = \mu_i + \nu_{i,T+S} \tag{12.38}$$

and $w = \sigma_\mu^2(\ell_i \otimes \iota_T)$ where ℓ_i is the i-th column of I_N, i.e., ℓ_i is a vector that has 1 in the i-th position and zero elsewhere. In this case

$$w'\Omega^{-1} = \sigma_\mu^2(\ell'_i \otimes \iota'_T)\left[\frac{1}{\sigma_1^2}P + \frac{1}{\sigma_\nu^2}Q\right] = \frac{\sigma_\mu^2}{\sigma_1^2}(\ell'_i \otimes \iota'_T) \tag{12.39}$$

since $(\ell'_i \otimes \iota'_T)\,P = (\ell'_i \otimes \iota'_T)$ and $(\ell'_i \otimes \iota'_T)Q = 0$. The typical element of $w'\Omega^{-1}\widehat{u}_{GLS}$ is $(T\sigma_\mu^2/\sigma_1^2)\widehat{\bar{u}}_{i.,GLS}$ where $\widehat{\bar{u}}_{i.,GLS} = \sum_{t=1}^{T}\widehat{u}_{it,GLS}/T$. Therefore, in (12.37), the BLUP for $y_{i,T+S}$ corrects the GLS prediction by a fraction of the mean of the GLS residuals corresponding to that i-th individual. This predictor was considered by Wansbeek and Kapteyn (1978) and Taub (1979).

12.4 Empirical Example

Baltagi and Griffin (1983) considered the following gasoline demand equation:

$$\log\frac{Gas}{Car} = \alpha + \beta_1\log\frac{Y}{N} + \beta_2\log\frac{P_{MG}}{P_{GDP}} + \beta_3\log\frac{Car}{N} + u \tag{12.40}$$

where Gas/Car is motor gasoline consumption per auto, Y/N is real income per capita, P_{MG}/P_{GDP} is real motor gasoline price and Car/N denotes the stock of cars per capita. This panel consists of annual observations across eighteen OECD countries, covering the period 1960-1978. The data for this example are provided on the Springer web site as GASOLINE.DAT. Table 12.1

gives the Stata output for the Within estimator using $xtreg, fe$. This is the regression described in (12.5) and computed as in (12.9). The Within estimator gives a low price elasticity for gasoline demand of -.322. The F-statistic for the significance of the country effects described in (12.14) yields an observed value of 83.96. This is distributed under the null as an $F(17, 321)$ and is statistically significant. This F-statistic is printed by Stata below the fixed effects output. In EViews, one invokes the test for redundant effects after running the fixed effects regression.

Table 12.1 Fixed Effects Estimator – Gasoline Demand Data

	Coef.	Std. Err.	T	$P > \|t\|$	[95% Conf. Interval]	
$\log(Y/N)$	0.6622498	0.073386	9.02	0.000	0.5178715	0.8066282
$\log(P_{MG}/P_{GDP})$	–0.3217025	0.0440992	–7.29	0.000	–0.4084626	–0.2349425
$\log(Car/N)$	–0.6404829	0.0296788	–21.58	0.000	–0.6988725	–0.5820933
Constant	2.40267	0.2253094	10.66	0.000	1.959401	2.84594
sigma_u	0.34841289					
sigma_e	0.09233034					
Rho	0.93438173	(fraction of variance due to u_i)				

Table 12.2 gives the Stata output for the Between estimator using $xtreg, be$.This is based on the regression given in (12.24). The Between estimator yields a high price elasticity of gasoline demand of -.964. These results were also verified using TSP.

Table 12.2 Between Estimator – Gasoline Demand Data

	Coef.	Std. Err.	T	$P > \|t\|$	[95% Conf. Interval]	
$\log(Y/N)$	0.9675763	0.1556662	6.22	0.000	0.6337055	1.301447
$\log(P_{MG}/P_{GDP})$	–0.9635503	0.1329214	–7.25	0.000	–1.248638	–0.6784622
$\log(Car/N)$	–0.795299	0.0824742	–9.64	0.000	–0.9721887	–0.6184094
Constant	2.54163	0.5267845	4.82	0.000	1.411789	3.67147

Table 12.3 gives the Stata output for the random effect model using $xtreg, re$. This is the Swamy and Arora (1972) estimator which yields a price elasticity of -.420. This is closer to the Within estimator than the Between estimator.

Table 12.3 Random Effects Estimator – Gasoline Demand Data

	Coef.	Std. Err.	T	$P > \|t\|$	[95% Conf. Interval]	
$\log(Y/N)$	0.5549858	0.0591282	9.39	0.000	0.4390967	0.6708749
$\log(P_{MG}/P_{GDP})$	–0.4203893	0.0399781	–10.52	0.000	–0.498745	–0.3420336
$\log(Car/N)$	–0.6068402	0.025515	–23.78	0.000	–0.6568487	–0.5568316
Constant	1.996699	0.184326	10.83	0.000	1.635427	2.357971
sigma_u	0.19554468					
sigma_e	0.09233034					
Rho	0.81769	(fraction of variance due to u_i)				

Table 12.4 Gasoline Demand Data. One-way Error Component Results

	β_1	β_2	β_3	ρ
OLS	0.890	−0.892	−0.763	0
	(0.036)*	(0.030)*	(0.019)*	
WALHUS	0.545	−0.447	−0.605	0.75
	(0.066)	(0.046)	(0.029)	
AMEMIYA	0.602	−0.366	−0.621	0.93
	(0.066)	(0.042)	(0.029)	
SWAR	0.555	−0.402	−0.607	0.82
	(0.059)	(0.042)	(0.026)	
IMLE	0.588	−0.378	−0.616	0.91
	(0.066)	(0.046)	(0.029)	

* These are biased standard errors when the true model has error component disturbances (see Moulton, 1986).
Source: Baltagi and Griffin (1983). Reproduced by permission of Elsevier Science Publishers B.V. (North-Holland).

Table 12.5 Gasoline Demand Data. Wallace and Hussain (1969) Estimator

Dependent Variable: GAS
Method: Panel EGLS (Cross-section random effects)

Sample: 1960 1978
Periods included: 19
Cross-sections included: 18
Total panel (balanced) observations: 342
Wallace and Hussain estimator of component variances

	Coefficient	Std. Error	t-Statistic	Prob.
C	1.938318	0.201817	9.604333	0.0000
$\log(Y/N)$	0.545202	0.065555	8.316682	0.0000
$\log(P_{MG}/P_{GDP})$	−0.447490	0.045763	−9.778438	0.0000
$\log(Car/N)$	−0.605086	0.028838	−20.98191	0.0000

Effects Specification		
	S.D.	Rho
Cross-section random	0.196715	0.7508
Idiosyncratic random	0.113320	0.2492

Table 12.4 gives the parameter estimates for OLS and three feasible GLS estimates of the slope coefficients along with their standard errors, and the corresponding estimate of ρ defined in (12.16). These were obtained using EViews by invoking the random effects estimation on the individual effects and choosing the estimation method from the options menu. Breusch's (1987) iterative maximum likelihood was computed using Stata$(xtreg, mle)$ and TSP.

Table 12.5 gives the EViews output for the Wallace and Hussain (1969) random effects estimator, while Table 12.6 gives the EViews output for the Amemiya (1971) random effects estimator. Note that EViews calls the Amemiya estimator Wansbeek and Kapteyn (1989) since the latter paper generalizes this method to deal with unbalanced panels with missing observations, see Baltagi (2005) for details. Table 12.6 gives the Stata maximum likelihood output.

Table 12.6 Gasoline Demand Data. Wansbeek and Kapteyn (1989) Estimator

Dependent Variable: GAS
Method: Panel EGLS (Cross-section random effects)

Sample: 1960 1978
Periods included: 19
Cross-sections included: 18
Total panel (balanced) observations: 342
Wallace and Hussain estimator of component variances

	Coefficient	Std. Error	t-Statistic	Prob.
C	2.188322	0.216372	10.11372	0.0000
$\log(Y/N)$	0.601969	0.065876	9.137941	0.0000
$\log(P_{MG}/P_{GDP})$	−0.365500	0.041620	−8.781832	0.0000
$\log(Car/N)$	−0.620725	0.027356	−22.69053	0.0000
	Effects Specification			
			S.D.	Rho
Cross-section random			0.343826	0.9327
Idiosyncratic random			0.092330	0.0673

Table 12.7 Gasoline Demand Data. Random Effects Maximum Likelihood Estimator

. xtreg c y p car,mle

Random-effects ML regression		Number of obs =	342
Group variable (i): coun		Number of groups =	18
Random effects u_i ∼ Gaussian		Obs per group: min =	19
		avg =	19.0
		max =	19
		LR chi2(3) =	609.75
Log likelihood = 282.47697		Prob > chi2 =	0.0000

c	Coef.	Std. Err.	z	$P>\|z\|$	[95% Conf.	Interval]
$\log(Y/N)$	.5881334	.0659581	8.92	0.000	.4588578	.717409
$\log(P_{MG}/P_{GDP})$	−.3780466	.0440663	−8.58	0.000	−.464415	−.2916782
$\log(Car/N)$	−.6163722	.0272054	−22.66	0.000	−.6696938	−.5630506
_cons	2.136168	.2156039	9.91	0.000	1.713593	2.558744
sigma_u	.2922939	.0545496			.2027512	.4213821
sigma_e	.0922537	.0036482			.0853734	.0996885
rho	.9094086	.0317608			.8303747	.9571561

Likelihood-ratio test of sigma_u = 0: chibar2(01)= 463.97 Prob >= chibar2 = 0.000

12.5 Testing in a Pooled Model

(1) The Chow-Test

Before pooling the data one may be concerned whether the data is poolable. This hypothesis is also known as the stability of the regression equation across firms or across time. It can be formulated in terms of an unrestricted model which involves a separate regression equation for each firm

$$y_i = Z_i\delta_i + u_i \quad \text{for } i = 1, 2, \ldots, N \tag{12.41}$$

where $y_i' = (y_{i1}, \ldots, y_{iT})$, $Z_i = [\iota_T, X_i]$ and X_i is $(T \times K)$. δ_i' is $1 \times (K+1)$ and u_i is $T \times 1$. The important thing to notice is that δ_i is different for every regional equation. We want to test the hypothesis H_0; δ_i=δ for all i, versus H_1; $\delta_i \neq \delta$ for some i. Under H_0 we can write the restricted model given in (12.41) as:

$$y = Z\delta + u \tag{12.42}$$

where $Z' = (Z_1', Z_2', \ldots, Z_N')$ and $u' = (u_1', u_2', \ldots, u_N')$. The unrestricted model can also be written as

$$y = \begin{pmatrix} Z_1 & 0 & \ldots & 0 \\ 0 & Z_2 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & Z_N \end{pmatrix} \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_N \end{pmatrix} + u = Z^*\delta^* + u \tag{12.43}$$

where $\delta^{*\prime} = (\delta_1', \delta_2', \ldots, \delta_N')$ and $Z = Z^*I^*$ with $I^* = (\iota_N \otimes I_{K'})$, an $NK' \times K'$ matrix, with $K' = K + 1$. Hence the variables in Z are all linear combinations of the variables in Z^*. Under the assumption that $u \sim N(0, \sigma^2 I_{NT})$, the MVU estimator for δ in equation (12.42) is

$$\widehat{\delta}_{OLS} = \widehat{\delta}_{MLE} = (Z'Z)^{-1}Z'y \tag{12.44}$$

and therefore

$$y = Z\widehat{\delta}_{OLS} + e \tag{12.45}$$

implying that $e = (I_{NT} - Z(Z'Z)^{-1}Z')y = My = M(Z\delta + u) = Mu$ since $MZ = 0$. Similarly, under the alternative, the MVU for δ_i is given by

$$\widehat{\delta}_{i,OLS} = \widehat{\delta}_{i,MLE} = (Z_i'Z_i)^{-1}Z_i'y_i \tag{12.46}$$

and therefore

$$y_i = Z_i\widehat{\delta}_{i,OLS} + e_i \tag{12.47}$$

implying that $e_i = (I_T - Z_i(Z_i'Z_i)^{-1}Z_i')y_i = M_iy_i = M_i(Z_i\delta_i + u_i) = M_iu_i$ since $M_iZ_i = 0$, and this is true for $i = 1, 2, \ldots, N$. Also, let

$$M^* = I_{NT} - Z^*(Z^{*\prime}Z^*)^{-1}Z^{*\prime} = \begin{pmatrix} M_1 & 0 & \ldots & 0 \\ 0 & M_2 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & M_N \end{pmatrix}$$

One can easily deduce that $y = Z^*\hat{\delta}^* + e^*$ with $e^* = M^*y = M^*u$ and $\hat{\delta}^* = (Z^{*\prime}Z^*)^{-1}Z^{*\prime}y$. Note that both M and M^* are symmetric and idempotent with $MM^* = M^*$. This easily follows since

$$Z(Z'Z)^{-1}Z'Z^*(Z^{*\prime}Z^*)^{-1}Z^{*\prime} = Z(Z'Z)^{-1}I^{*\prime}Z^{*\prime}Z^*(Z^{*\prime}Z^*)^{-1}Z^{*\prime} = Z(Z'Z)^{-1}Z'$$

This uses the fact that $Z = Z^*I^*$. Now, $e'e - e^{*\prime}e^* = u'(M - M^*)u$ and $e^{*\prime}e^* = u'M^*u$ are independent since $(M - M^*)M^* = 0$. Also, both quadratic forms when divided by σ^2 are distributed as χ^2 since $(M - M^*)$ and M^* are idempotent, see Judge et al. (1985). Dividing these quadratic forms by their respective degrees of freedom, and taking their ratio leads to the following test statistic:

$$\begin{aligned} F_{obs} &= \frac{(e'e - e^{*\prime}e^*)/(\text{tr}(M) - \text{tr}(M^*))}{e^{*\prime}e^*/\text{tr}(M^*)} \qquad (12.48) \\ &= \frac{(e'e - e_1'e_1 - e_2'e_2 - .. - e_N'e_N)/(N-1)K'}{(e_1'e_1 + e_2'e_2 + .. + e_N'e_N)/N(T-K')} \end{aligned}$$

Under H_0, F_{obs} is distributed as an $F((N-1)K', N(T-K'))$, see lemma 2.2 of Fisher (1970). This is exactly the Chow's (1960) test extended to the case of N linear regressions.

The URSS in this case is the sum of the N residual sum of squares obtained by applying OLS to (12.41), i.e., on each firm equation separately. The RRSS is simply the RSS from OLS performed on the pooled regression given by (12.42). In this case, there are $(N-1)K'$ restrictions and the URSS has $N(T-K')$ degrees of freedom. Similarly, one can test the stability of the regression across time. In this case, the degrees of freedom are $(T-1)K'$ and $N(T-K')$ respectively. Both tests target the whole set of regression coefficients including the constant. If the LSDV model is suspected to be the proper specification, then the intercepts are allowed to vary but the slopes remain the same. To test the stability of the slopes only, the same Chow-test can be utilized, however the RRSS is now that of the LSDV regression with firm (or time) dummies only. The number of restrictions becomes $(N-1)K$ for testing the stability of the slopes across firms and $(T-1)K$ for testing their stability across time.

The Chow-test however is proper under spherical disturbances, and if that hypothesis is not correct it will lead to improper inference. Baltagi (1981) showed that if the true specification of the disturbances is an error components structure then the Chow-test tend to reject poolability too often when in fact it is true. However, a generalization of the Chow-test which takes care of the general variance-covariance matrix is available in Zellner (1962). This is exactly the test of the null hypothesis H_0; $R\beta = r$ when Ω is that of the error components specification, see Chapter 9. Baltagi (1981) shows that this test performs well in Monte Carlo experiments. In this case, all we need to do is transform our model (under both the null and alternative hypotheses) such that the transformed disturbances have a variance of $\sigma^2 I_{NT}$, then apply the Chow-test on the transformed model. The later step is legitimate because the transformed disturbances have homoskedastic variances and the usual Chow-test is legitimate. Given $\Omega = \sigma^2\Sigma$, we premultiply the restricted model given in (12.42) by $\Sigma^{-1/2}$ and we call $\Sigma^{-1/2}y = \dot{y}$, $\Sigma^{-1/2}Z = \dot{Z}$ and $\Sigma^{-1/2}u = \dot{u}$. Hence

$$\dot{y} = \dot{Z}\delta + \dot{u} \qquad (12.49)$$

with $E(\dot{u}\dot{u}') = \Sigma^{-1/2}E(uu')\Sigma^{-1/2\prime} = \sigma^2 I_{NT}$. Similarly, we premultiply the unrestricted model given in (12.43) by $\Sigma^{-1/2}$ and we call $\Sigma^{-1/2}Z^* = \dot{Z}^*$. Therefore

$$\dot{y} = \dot{Z}^* \delta^* + \dot{u} \tag{12.50}$$

with $E(\dot{u}\dot{u}') = \sigma^2 I_{NT}$.

At this stage, we can test H_0; $\delta_i = \delta$ for every $i = 1, 2, \ldots, N$, simply by using the Chow-statistic, only now on the transformed models (12.49) and (12.50) since they satisfy $\dot{u} \sim N(0, \sigma^2 I_{NT})$. Note that $\dot{Z} = \dot{Z}^* I^*$ which is simply obtained from $Z = Z^* I^*$ by premultiplying by $\Sigma^{-1/2}$. Defining $\dot{M} = I_{NT} - \dot{Z}(\dot{Z}'\dot{Z})^{-1}\dot{Z}'$, and $\dot{M}^* = I_{NT} - \dot{Z}^*(\dot{Z}^{*\prime}\dot{Z}^*)^{-1}\dot{Z}^{*\prime}$, it is easy to show that $\dot{M}$ and $\dot{M}^*$ are both symmetric, idempotent and such that $\dot{M}\dot{M}^* = \dot{M}^*$. Once again the conditions for lemma 2.2 of Fisher (1970) are satisfied, and the test-statistic

$$\dot{F}_{obs} = \frac{(\dot{e}'\dot{e} - \dot{e}^{*\prime}\dot{e}^*)/(\mathrm{tr}(\dot{M}) - \mathrm{tr}(\dot{M}^*))}{\dot{e}^{*\prime}\dot{e}^*/\mathrm{tr}(\dot{M}^*)} \sim F((N-1)K',\, N(T-K')) \tag{12.51}$$

where $\dot{e} = \dot{y} - \dot{Z}\widehat{\dot{\delta}}_{OLS}$ and $\widehat{\dot{\delta}}_{OLS} = (\dot{Z}'\dot{Z})^{-1}\dot{Z}'\dot{y}$ implying that $\dot{e} = \dot{M}\dot{y} = \dot{M}\dot{u}$. Similarly, $\dot{e}^* = \dot{y} - \dot{Z}^*\widehat{\dot{\delta}}^*_{OLS}$ and $\widehat{\dot{\delta}}^*_{OLS} = (\dot{Z}^{*\prime}\dot{Z}^*)^{-1}\dot{Z}^{*\prime}\dot{y}$ implying that $\dot{e}^* = \dot{M}^*\dot{y} = \dot{M}^*\dot{u}$. This is the Chow-test after premultiplying the model by $\Sigma^{-1/2}$ or simply applying the Fuller and Battese (1974) transformation. See Baltagi (2005) for details.

For the gasoline data in Baltagi and Griffin (1983), Chow's test for poolability across countries yields an observed F-statistic of 129.38 and is distributed as $F(68, 270)$ under H_0; $\delta_i = \delta$ for $i = 1, \ldots, N$. This tests the stability of four time-series regression coefficients across 18 countries. The unrestricted SSE is based upon 18 OLS time-series regressions, one for each country. For the stability of the slope coefficients only, H_0; $\beta_i = \beta$, an observed F-value of 27.33 is obtained which is distributed as $F(51, 270)$ under the null. Chow's test for poolability across time yields an F-value of 0.276 which is distributed as $F(72, 266)$ under H_0; $\delta_t = \delta$ for $t = 1, \ldots, T$. This tests the stability of four cross-section regression coefficients across 19 time periods. The unrestricted SSE is based upon 19 OLS cross-section regressions, one for each year. This does not reject poolability across time-periods. The test for poolability across countries, allowing for a one-way error components model yields an F-value of 21.64 which is distributed as $F(68, 270)$ under H_0; $\delta_i = \delta$ for $i = 1, \ldots, N$. The test for poolability across time yields an F-value of 1.66 which is distributed as $F(72, 266)$ under H_0; $\delta_t = \delta$ for $t = 1, \ldots, T$. This rejects H_0 at the 5% level.

(2) The Breusch-Pagan Test

Next, we look at a Lagrange Multiplier test developed by Breusch and Pagan (1980), which tests whether H_0; $\sigma^2_\mu = 0$. The test statistic is given by

$$LM = (NT/2(T-1)) \left[(\textstyle\sum_{i=1}^N e_{i.}^2 / \sum_{i=1}^N \sum_{t=1}^T e_{it}^2) - 1 \right]^2 \tag{12.52}$$

where e_{it} denotes the OLS residuals on the pooled model, $e_{i.}$ denote their sum over t, respectively. Under the null hypothesis H_0 this LM statistic is distributed as a χ^2_1. For the gasoline data in Baltagi and Griffin (1983), the Breusch and Pagan LM test yields an LM statistic of 1465.6. This is obtained using the Stata command *xtest0* after estimating the model with random effects. This is significant and rejects the null hypothesis. The corresponding likelihood ratio test assuming Normal disturbances is also reported by Stata maximum likelihood output for the random effects model. This yields an LR statistic of 463.97 which is asymptotically distributed as χ^2_1 under the null hypothesis H_0 and is also significant.

One problem with the Breusch-Pagan test is that it assumes that the alternative hypothesis is two-sided when we know that $\sigma_\mu^2 > 0$. A one-sided version of this test is given by Honda (1985):

$$HO = \sqrt{\frac{NT}{2(T-1)}}\left[\frac{e'(I_N \otimes J_T)e}{e'e} - 1\right] \overset{H_0}{\to} N(0,1) \tag{12.53}$$

where e denotes the vector of OLS residuals. Note that the square of this $N(0,1)$ statistic is the Breusch and Pagan (1980) LM test-statistic. Honda (1985) finds that this test statistic is *uniformly most powerful* and robust to non-normality. However, Moulton and Randolph (1989) showed that the asymptotic $N(0,1)$ approximation for this one-sided LM statistic can be poor even in large samples. They suggest an alternative Standardized Lagrange Multiplier (SLM) test whose asymptotic critical values are generally closer to the exact critical values than those of the LM test. This SLM test statistic centers and scales the one-sided LM statistic so that its mean is zero and its variance is one.

$$SLM = \frac{HO - E(HO)}{\sqrt{\text{var}(HO)}} = \frac{d - E(d)}{\sqrt{\text{var}(d)}} \tag{12.54}$$

where $d = e'De/e'e$ and $D = (I_N \otimes J_T)$. Using the results on moments of quadratic forms in regression residuals, see for e.g., Evans and King (1985), we get

$$E(d) = \text{tr}(D\bar{P}_Z)/p$$

and

$$\text{var}(d) = 2\{p\ \text{tr}(D\bar{P}_Z)^2 - [\text{tr}(D\bar{P}_Z)]^2\}/p^2(p+2) \tag{12.55}$$

where $p = n - (K+1)$ and $\bar{P}_Z = I_n - Z(Z'Z)^{-1}Z'$. Under the null hypothesis, SLM has an asymptotic $N(0,1)$ distribution.

(3) The Hausman-Test

A critical assumption in the error components regression model is that $E(u_{it}/X_{it}) = 0$. This is important given that the disturbances contain individual effects (the μ_i's) which are unobserved and may be correlated with the X_{it}'s. For example, in an earnings equation these μ_i's may denote unobservable ability of the individual and this may be correlated with the schooling variable included on the right hand side of this equation. In this case, $E(u_{it}/X_{it}) \neq 0$ and the GLS estimator $\widehat{\beta}_{GLS}$ becomes biased and inconsistent for β. However, the within transformation wipes out these μ_i's and leaves the Within estimator $\widetilde{\beta}_{Within}$ unbiased and consistent for β. Hausman (1978) suggests comparing $\widehat{\beta}_{GLS}$ and $\widetilde{\beta}_{Within}$, both of which are consistent under the null hypothesis H_0; $E(u_{it}/X_{it}) = 0$, but which will have different probability limits if H_0 is not true. In fact, $\widetilde{\beta}_{Within}$ is consistent whether H_0 is true or not, while $\widehat{\beta}_{GLS}$ is BLUE, consistent and asymptotically efficient under H_0, but is inconsistent when H_0 is false. A natural test statistic would be based on $\widehat{q} = \widehat{\beta}_{GLS} - \widetilde{\beta}_{Within}$. Under H_0, plim $\widehat{q} = 0$, and $\text{cov}(\widehat{q}, \widehat{\beta}_{GLS}) = 0$.

Using the fact that $\widehat{\beta}_{GLS} - \beta = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}u$ and $\widetilde{\beta}_{Within} - \beta = (X'QX)^{-1}X'Qu$, one gets $E(\widehat{q}) = 0$ and

$$\begin{aligned}\text{cov}(\widehat{\beta}_{GLS}, \widehat{q}) &= \text{var}(\widehat{\beta}_{GLS}) - \text{cov}(\widehat{\beta}_{GLS}, \widetilde{\beta}_{Within}) \\ &= (X'\Omega^{-1}X)^{-1} - (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}E(uu')QX(X'QX)^{-1} = 0\end{aligned}$$

Using the fact that $\widetilde{\beta}_{Within} = \widehat{\beta}_{GLS} - \widehat{q}$, one gets

$$\text{var}(\widetilde{\beta}_{Within}) = \text{var}(\widehat{\beta}_{GLS}) + \text{var}(\widehat{q}),$$

since $\text{cov}(\widehat{\beta}_{GLS}, \widehat{q}) = 0$. Therefore,

$$\text{var}(\widehat{q}) = \text{var}(\widetilde{\beta}_{Within}) - \text{var}(\widehat{\beta}_{GLS}) = \sigma^2_\nu (X'QX)^{-1} - (X'\Omega^{-1}X)^{-1} \tag{12.56}$$

Hence, the Hausman test statistic is given by

$$m = \widehat{q}'[\text{var}(\widehat{q})]^{-1}\widehat{q} \tag{12.57}$$

and under H_0 is asymptotically distributed as χ^2_K, where K denotes the dimension of slope vector β. In order to make this test operational, Ω is replaced by a consistent estimator $\widehat{\Omega}$, and GLS by its corresponding FGLS. An alternative asymptotically equivalent test can be obtained from the augmented regression

$$y^* = X^*\beta + \widetilde{X}\gamma + w \tag{12.58}$$

where $y^* = \sigma_\nu \Omega^{-1/2} y$, $X^* = \sigma_\nu \Omega^{-1/2} X$ and $\widetilde{X} = QX$. Hausman's test is now equivalent to testing whether $\gamma = 0$. This is a standard Wald test for the omission of the variables $\widetilde{X}$ from (12.58).

This test was generalized by Arellano (1993) to make it robust to heteroskedasticity and autocorrelation of arbitrary forms. In fact, if either heteroskedasticity or serial correlation is present, the variances of the Within and GLS estimators are not valid and the corresponding Hausman test statistic is inappropriate. For the Baltagi and Griffin (1983) gasoline data, the Hausman test statistic based on the difference between the Within estimator and that of feasible GLS based on Swamy and Arora (1972) yields a χ^2_3 value of $m = 306.1$ which rejects the null hypothesis. This is obtained using the Stata command *hausman*.

12.6 Dynamic Panel Data Models

The dynamic error components regression is characterized by the presence of a lagged dependent variable among the regressors, i.e.,

$$y_{it} = \delta y_{i,t-1} + x'_{it}\beta + \mu_i + \nu_{it}, \quad i = 1, \ldots, N; \quad t = 1, \ldots, T \tag{12.59}$$

where δ is a scalar, x'_{it} is $1 \times K$ and β is $K \times 1$. This model has been extensively studied by Anderson and Hsiao (1982). Since y_{it} is a function of μ_i, $y_{i,t-1}$ is also a function of μ_i. Therefore, $y_{i,t-1}$, a right hand regressor in (12.59), is correlated with the error term. This renders the OLS estimator biased and inconsistent even if the ν_{it}'s are not serially correlated. For the FE estimator, the within transformation wipes out the μ_i's, but $\tilde{y}_{i,t-1}$ will still be correlated with $\tilde{\nu}_{it}$ even if the ν_{it}'s are not serially correlated. In fact, the Within estimator will be biased of $O(1/T)$ and its consistency will depend upon T being large, see Nickell (1981). An alternative transformation that wipes out the individual effects, yet does not create the above problem is the first difference (FD) transformation. In fact, Anderson and Hsiao (1982) suggested first differencing the model to get rid of the μ_i's and then using $\Delta y_{i,t-2} = (y_{i,t-2} - y_{i,t-3})$ or simply $y_{i,t-2}$ as an instrument for $\Delta y_{i,t-1} = (y_{i,t-1} - y_{i,t-2})$. These instruments will not be

correlated with $\Delta\nu_{it} = \nu_{i,t} - \nu_{i,t-1}$, as long as the ν_{it}'s themselves are not serially correlated. This instrumental variable (IV) estimation method leads to consistent but not necessarily efficient estimates of the parameters in the model. This is because it does not make use of all the available moment conditions, see Ahn and Schmidt (1995), and it does not take into account the differenced structure on the residual disturbances ($\Delta\nu_{it}$). Arellano (1989) finds that for simple dynamic error components models the estimator that uses differences $\Delta y_{i,t-2}$ rather than levels $y_{i,t-2}$ for instruments has a singularity point and very large variances over a significant range of parameter values. In contrast, the estimator that uses instruments in levels, i.e., $y_{i,t-2}$, has no singularities and much smaller variances and is therefore recommended. Additional instruments can be obtained in a dynamic panel data model if one utilizes the orthogonality conditions that exist between lagged values of y_{it} and the disturbances ν_{it}.

Let us illustrate this with the simple autoregressive model with no regressors:

$$y_{it} = \delta y_{i,t-1} + u_{it} \quad i = 1, \ldots, N \quad t = 1, \ldots, T \tag{12.60}$$

where $u_{it} = \mu_i + \nu_{it}$ with $\mu_i \sim \text{IID}(0, \sigma_\mu^2)$ and $\nu_{it} \sim \text{IID}(0, \sigma_\nu^2)$, independent of each other and among themselves. In order to get a consistent estimate of δ as $N \to \infty$ with T fixed, we first difference (12.60) to eliminate the individual effects

$$y_{it} - y_{i,t-1} = \delta(y_{i,t-1} - y_{i,t-2}) + (\nu_{it} - \nu_{i,t-1}) \tag{12.61}$$

and note that $(\nu_{it} - \nu_{i,t-1})$ is MA(1) with unit root. For the first period we observe this relationship, i.e., $t = 3$, we have

$$y_{i3} - y_{i2} = \delta(y_{i2} - y_{i1}) + (\nu_{i3} - \nu_{i2})$$

In this case, y_{i1} is a valid instrument, since it is highly correlated with $(y_{i2} - y_{i1})$ and not correlated with $(\nu_{i3} - \nu_{i2})$ as long as the ν_{it} are not serially correlated. But note what happens for $t = 4$, the second period we observe (12.61):

$$y_{i4} - y_{i3} = \delta(y_{i3} - y_{i2}) + (\nu_{i4} - \nu_{i3})$$

In this case, y_{i2} as well as y_{i1} are valid instruments for $(y_{i3} - y_{i2})$, since both y_{i2} and y_{i1} are not correlated with $(\nu_{i4} - \nu_{i3})$. One can continue in this fashion, adding an extra valid instrument with each forward period, so that for period T, the set of valid instruments becomes $(y_{i1}, y_{i2}, \ldots, y_{i,T-2})$.

This instrumental variable procedure still does not account for the differenced error term in (12.61). In fact,

$$E(\Delta\nu_i \, \Delta\nu_i') = \sigma_\nu^2 G \tag{12.62}$$

where $\Delta\nu_i' = (\nu_{i3} - \nu_{i2}, \ldots, \nu_{iT} - \nu_{i,T-1})$ and

$$G = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 2 \end{pmatrix}$$

is $(T-2)\times(T-2)$, since $\Delta\nu_i$ is MA(1) with unit root. Define

$$W_i = \begin{bmatrix} [y_{i1}] & & & 0 \\ & [y_{i1}, y_{i2}] & & \\ & & \ddots & \\ 0 & & & [y_{i1}, \ldots, y_{i,T-2}] \end{bmatrix} \tag{12.63}$$

Then, the matrix of instruments is $W = [W_1', \ldots, W_N']'$ and the moment equations described above are given by $E(W_i'\Delta\nu_i) = 0$. Premultiplying the differenced equation (12.61) in vector form by W', one gets

$$W'\Delta y = W'(\Delta y_{-1})\delta + W'\Delta\nu \tag{12.64}$$

Performing GLS on (12.64) one gets the Arellano and Bond (1991) preliminary one-step consistent estimator

$$\begin{aligned} \widehat{\delta}_1 &= [(\Delta y_{-1})'W(W'(I_N\otimes G)W)^{-1}W'(\Delta y_{-1})]^{-1} \\ &\quad \times[(\Delta y_{-1})'W(W'(I_N\otimes G)W)^{-1}W'(\Delta y)] \end{aligned} \tag{12.65}$$

The optimal generalized method of moments (GMM) estimator of δ_1 à la Hansen (1982) for $N\to\infty$ and T fixed using only the above moment restrictions yields the same expression as in (12.65) except that

$$W'(I_N\otimes G)W = \sum_{i=1}^{N} W_i'GW_i$$

is replaced by

$$V_N = \sum_{i=1}^{N} W_i'(\Delta\nu_i)(\Delta\nu_i)'W_i$$

This GMM estimator requires no knowledge concerning the initial conditions or the distributions of ν_i and μ_i. To operationalize this estimator, $\Delta\nu$ is replaced by differenced residuals obtained from the preliminary consistent estimator $\widehat{\delta}_1$. The resulting estimator is the two-step Arellano and Bond (1991) GMM estimator:

$$\widehat{\delta}_2 = [(\Delta y_{-1})'W\widehat{V}_N^{-1}W'(\Delta y_{-1})]^{-1}[(\Delta y_{-1})'W\widehat{V}_N^{-1}W'(\Delta y)] \tag{12.66}$$

A consistent estimate of the asymptotic $\mathrm{var}(\widehat{\delta}_2)$ is given by the first term in (12.66),

$$\widehat{\mathrm{var}}(\widehat{\delta}_2) = [(\Delta y_{-1})'W\widehat{V}_N^{-1}W'(\Delta y_{-1})]^{-1} \tag{12.67}$$

Note that $\widehat{\delta}_1$ and $\widehat{\delta}_2$ are asymptotically equivalent if the ν_{it} are $\mathrm{IID}(0,\sigma_\nu^2)$.

If there are additional *strictly exogenous* regressors x_{it} as in (12.59) with $E(x_{it}\nu_{is}) = 0$ for all $t,s = 1,2,\ldots,T$, but where all the x_{it} are correlated with μ_i, then all the x_{it} are valid instruments for the first differenced equation of (12.59). Therefore, $[x_{i1}', x_{i2}', \ldots, x_{iT}']$ should be added to each diagonal element of W_i in (12.63). In this case, (12.64) becomes

$$W'\Delta y = W'(\Delta y_{-1})\delta + W'(\Delta X)\beta + W'\Delta\nu$$

where ΔX is the stacked $N(T-2) \times K$ matrix of observations on Δx_{it}. One- and two-step estimators of (δ, β')can be obtained from

$$\begin{pmatrix} \widehat{\delta} \\ \widehat{\beta} \end{pmatrix} = ([\Delta y_{-1}, \Delta X]' W \widehat{V}_N^{-1} W' [\Delta y_{-1}, \Delta X])^{-1} ([\Delta y_{-1}, \Delta X]' W \widehat{V}_N^{-1} W' \Delta y) \tag{12.68}$$

as in (12.65) and (12.66).

Arellano and Bond (1991) suggest Sargan's (1958) test of over-identifying restrictions given by

$$m = (\Delta \widehat{\nu})' W \left[\sum_{i=1}^{N} W_i' (\Delta \widehat{\nu}_i)(\Delta \widehat{\nu}_i)' W_i \right]^{-1} W' (\Delta \widehat{\nu}) \sim \chi^2_{p-K-1}$$

where p refers to the number of columns of W and $\Delta \widehat{\nu}$ denote the residuals from a two-step estimation given in (12.68).

To summarize, dynamic panel data estimation of equation (12.59) with individual fixed effects suffers from the Nickell (1981) bias. This disappears only if T tends to infinity. Alternatively, a GMM estimator was suggested by Arellano and Bond (1991) which basically differences the model to get rid of the individual specific effects and along with it any time invariant regressor. This also gets rid of any endogeneity that may be due to the correlation of these individual effects and the right hand side regressors. The moment conditions utilize the orthogonality conditions between the differenced errors and lagged values of the dependent variable. This assumes that the original disturbances are serially uncorrelated. In fact, two diagnostics are computed using the Arellano and Bond GMM procedure to test for first order and second order serial correlation in the disturbances. One should reject the null of the absence of first order serial correlation and not reject the absence of second order serial correlation. A special feature of dynamic panel data GMM estimation is that the number of moment conditions increase with T. Therefore, a Sargan test is performed to test the over-identification restrictions. There is convincing evidence that too many moment conditions introduce bias while increasing efficiency. It is even suggested that a subset of these moment conditions be used to take advantage of the trade-off between the reduction in bias and the loss in efficiency, see Baltagi (2005) for details.

Arellano and Bond (1991) apply their GMM estimation and testing methods to a model of employment using a panel of 140 quoted UK companies for the period 1979-84. This is the benchmark data set used in Stata to obtain the one-step and two-step estimators described in (12.65) and (12.66) as well as the Sargan test for over-identification using the command (xtabond,twostep).The reader is asked to replicate their results in problem 22.

12.6.1 Empirical Illustration

Baltagi, Griffin and Xiong (2000) estimate a dynamic demand model for cigarettes based on panel data from 46 American states over 30 years 1963-1992. The estimated equation is

$$\ln C_{it} = \alpha + \beta_1 \ln C_{i,t-1} + \beta_2 \ln P_{i,t} + \beta_3 \ln Y_{it} + \beta_4 \ln Pn_{it} + u_{it} \tag{12.69}$$

where the subscript i denotes the ith state ($i = 1, \ldots, 46$), and the subscript t denotes the tth year ($t = 1, \ldots, 30$). C_{it} is real per capita sales of cigarettes by persons of smoking age (14

years and older). This is measured in packs of cigarettes per head. P_{it} is the average retail price of a pack of cigarettes measured in real terms. Y_{it} is real per capita disposable income. Pn_{it} denotes the minimum real price of cigarettes in any neighboring state. This last variable is a proxy for the casual smuggling effect across state borders. It acts as a substitute price attracting consumers from high-tax states like Massachusetts to cross over to New Hampshire where the tax is low. The disturbance term is specified as a two-way error component model:

$$u_{it} = \mu_i + \lambda_t + \nu_{it} \quad i = 1, \ldots, 46 \quad t = 1, \ldots, 30 \tag{12.70}$$

where μ_i denotes a state-specific effect, and λ_t denotes a year-specific effect. The time-period effects (the λ_t) are assumed fixed parameters to be estimated as coefficients of time dummies for each year in the sample. This can be justified given the numerous policy interventions as well as health warnings and Surgeon General's reports. For example:

(1) the imposition of warning labels by the Federal Trade Commission effective January 1965;

(2) the application of the Fairness Doctrine Act to cigarette advertising in June 1967, which subsidized antismoking messages from 1968 to 1970;

(3) the Congressional ban on broadcast advertising of cigarettes effective January 1971.

The μ_i are state-specific effects which can represent any state-specific characteristic including the following:

(1) States with Indian reservations like Montana, New Mexico and Arizona are among the biggest losers in tax revenues from non-Indians purchasing tax-exempt cigarettes from the reservations.

(2) Florida, Texas, Washington and Georgia are among the biggest losers of revenues due to the purchasing of cigarettes from tax-exempt military bases in these states.

(3) Utah, which has a high percentage of Mormon population (a religion which forbids smoking), has a per capita sales of cigarettes in 1988 of 55 packs, a little less than half the national average of 113 packs.

(4) Nevada, which is a highly touristic state, has a per capita sales of cigarettes of 142 packs in 1988, 29 more packs than the national average.

These state-specific effects may be assumed fixed, in which case one includes state dummy variables in equation (12.69). The resulting estimator is the Within estimator reported in Table 12.8. Comparing these estimates with OLS without state or time dummies, one can see that the coefficient of lagged consumption drops from 0.97 to 0.83 and the price elasticity goes up in absolute value from -0.09 to -0.30. The income elasticity switches sign from negative to positive going from -0.03 to 0.10.

The OLS and Within estimators do not take into account the endogeneity of the lagged dependent variable, and therefore 2SLS and Within-2SLS are performed. The instruments used are one lag on price, neighboring price and income. These give lower estimates of lagged consumption and higher own price elasticities in absolute value. The Arellano and Bond (1991) two-step estimator yields an estimate of lagged consumption of 0.70 and a price elasticity of -0.40, both of which are significant. Sargan's test for over-identification yields an observed value of 32.3. This is asymptotically distributed as χ^2_{27} and is not significant. This was obtained using the Stata command $(xtabond2, twostep)$ with the collapse option to reduce the number of moment conditions used for estimation.

Table 12.8 Dynamic Demand for Cigarettes: 1963-92*

	$lnC_{i,t-1}$	lnP_{it}	lnY_{it}	$lnPn_{it}$
OLS	0.97	−0.090	−0.03	0.024
	(157.7)	(6.2)	(5.1)	(1.8)
Within	0.83	−0.299	0.10	0.034
	(66.3)	(12.7)	(4.2)	(1.2)
2SLS	0.85	−0.205	−0.02	0.052
	(25.3)	(5.8)	(2.2)	(3.1)
Within-2SLS	0.60	−0.496	0.19	−0.016
	(17.0)	(13.0)	(6.4)	(0.5)
Arellano and Bond (two-step)	0.70	−0.396	0.13	−0.003
	(10.2)	(6.0)	(3.5)	(0.1)

* Numbers in parentheses are t-statistics.
Source: Some of the results in this Table are reported in Baltagi, Griffin and Xiong (2000).

12.7 Program Evaluation and Difference-in-Differences Estimator

Suppose we want to study the effect of job training programs on earnings. An ideal experiment would assign individuals randomly, by a flip of a coin, to training and non-training camps, and then compare their earnings, holding other factors constant. This is a necessary experiment before the approval of any drug. Patients are randomly assigned to receive the drug or a placebo and the drug is approved or disapproved depending on the difference in the outcome between these two groups. In this case, the FDA is concerned with the drug's safety and its effectiveness. However, we run into problems in setting this experiment. How can we hold other factors constant? Even twins which have been used in economic studies are not identical and may have different life experiences.

The individual's prior work experience will affect one's chances in getting a job after training. But as long as the individuals are randomly assigned, the distribution of work experience is the same in the treatment and control group, i.e., participation in the job training is independent of prior work experience. In this case, omitting previous work experience from the analysis will not cause omitted variable bias in the estimator of the effect of the training program on future employment. Stock and Watson (2003) discuss threats to the internal and external validity of such experiments. The former include: (i) failure to randomize, or (ii) to follow the treatment protocol. These failures can cause bias in estimating the effect of the treatment. The first can happen when individuals are assigned non-randomly to the treatment and non-treatment groups. The second can happen, for example, when some people in the training program do not show up for all training sessions; or when some people who are not supposed to be in the training program are allowed to attend some of these training sessions. Attrition caused by people dropping out of the experiment in either group can cause bias especially if the cause of attrition is related to their acquiring or not acquiring training. In addition, small samples, usually associated with expensive experiments, can affect the precision of the estimates. There can also be experimental effects, brought about by people trying harder simply because the worker being trained feels noticed or because the trainer has a stake in the success of the program. Stock and Watson (2003, p. 380) argue that "threats to external validity compromise the ability to generalize the results of the experiment to other populations and settings. Two such threats are when the experimental sample is not representative of the population of interest

and when the treatment being studied is not representative of the treatment that would be implemented more broadly."

They also warn about "general equilibrium effects" where, for example, turning a small, temporary experimental program into a widespread, permanent program might change the economic environment sufficiently that the results of the experiment cannot be generalized. For example, it could displace employer-provided training, thereby reducing the net benefits of the program.

12.7.1 The Difference-in-Differences Estimator

With panel data, observations on the same subjects before and after the training program allow us to estimate the effect of this program on earnings. In simple regression form, assuming the assignment to the training program is random, one regresses the change in earnings before and after training is completed on a dummy variable which takes the value 1 if the individual received training and zero if they did not. This regression computes the average change in earnings for the treatment group before and after the training program and subtracts that from the average change in earnings for the control group. One can include additional regressors which measure the individual characteristics prior to training. Examples are gender, race, education and age of the individual.

Card (1990) used a quasi-experiment to see whether immigration reduces wages. Taking advantage of the "Mariel boatlift" where a large number of Cuban immigrants entered Miami. Card (1990) used the difference-in-differences estimator, comparing the change in wages of low-skilled workers in Miami to the change in wages of similar workers in other comparable U.S. cities over the same period. Card concluded that the influx of Cuban immigrants had a negligible effect on wages of less-skilled workers.

Problems

1. (a) Premultiply (12.11) by Q and verify that the transformed equation reduces to (12.12). Show that the new disturbances $Q\nu$ have zero mean and variance-covariance matrix $\sigma_\nu^2 Q$. **Hint:** $QZ_\mu = 0$.

   (b) Show that the GLS estimator is the same as the OLS estimator on this transformed regression equation. **Hint:** Use one of the necessary and sufficient conditions for GLS to be equivalent to OLS given in Chapter 9.

   (c) Using the Frisch-Waugh-Lovell Theorem given in Chapter 7, show that the estimator derived in part (b) is the Within estimator and is given by $\widetilde{\beta} = (X'QX)^{-1}X'Qy$.

2. (a) Show that Ω given in (12.17) can be written as (12.18).

   (b) Show that P and Q are symmetric, idempotent, orthogonal and sum to the identity matrix.

   (c) For Ω^{-1} given by (12.19), verify that $\Omega\Omega^{-1} = \Omega^{-1}\Omega = I_{NT}$.

   (d) For $\Omega^{-1/2}$ given by (12.20), verify that $\Omega^{-1/2}\Omega^{-1/2} = \Omega^{-1}$.

3. Premultiply y by $\sigma_\nu \Omega^{-1/2}$ where $\Omega^{-1/2}$ is defined in (12.20) and show that the resulting y^* has a typical element $y_{it}^* = y_{it} - \theta \bar{y}_{i.}$, where the $\theta = 1 - \sigma_\nu/\sigma_1$ and $\sigma_1^2 = T\sigma_\mu^2 + \sigma_\nu^2$.

4. Using (12.21) and (12.22), show that $E(\widehat{\sigma}_1^2) = \sigma_1^2$ and $E(\widehat{\sigma}_\nu^2) = \sigma_\nu^2$. **Hint:** $E(u'Qu) = E\{\text{tr}(u'Qu)\} = E\{\text{tr}(uu'Q)\} = \text{tr}\{E(uu')Q\} = \text{tr}(\Omega Q)$.

5. (a) Show that $\widehat{\widehat{\sigma}}^2_\nu$ given in (12.23) is unbiased for σ^2_ν.

   (b) Show that $\widehat{\widehat{\sigma}}^2_1$ given in (12.26) is unbiased for σ^2_1.

6. (a) Perform OLS on the system of equations given in (12.27) and show that the resulting estimator is $\widehat{\delta}_{OLS} = (Z'Z)^{-1}Z'y$.

   (b) Perform GLS on this system of equations and show that the resulting estimator is $\widehat{\delta}_{GLS} = (Z'\Omega^{-1}Z)^{-1}Z'\Omega^{-1}y$ where Ω^{-1} is given in (12.19).

7. Using the $\text{var}(\widehat{\beta}_{GLS})$ expression below (12.30) and $\text{var}(\widetilde{\beta}_{Within}) = \sigma^2_\nu W^{-1}_{XX}$, show that

$$(\text{var}(\widehat{\beta}_{GLS}))^{-1} - (\text{var}(\widetilde{\beta}_{Within}))^{-1} = \phi^2 B_{XX}/\sigma^2_\nu$$

   which is positive semi-definite. Conclude that $\text{var}(\widetilde{\beta}_{Within}) - \text{var}(\widehat{\beta}_{GLS})$ is positive semi-definite.

8. (a) Using the concentrated likelihood function in (12.34), solve $\partial L_c/\partial\phi^2 = 0$ and verify (12.35).

   (b) Solve $\partial L_c/\partial\beta = 0$ and verify (12.36).

9. (a) For the predictor of $y_{i,T+S}$ given in (12.37), compute $E(u_{i,T+S}u_{it})$ for $t = 1, 2, \ldots, T$ and verify that $w = E(u_{i,T+S}u) = \sigma^2_\mu(\ell_i \otimes \iota_T)$ where ℓ_i is the i-th column of I_N.

   (b) Verify (12.39) by showing that $(\ell'_i \otimes \iota'_T)P = (\ell'_i \otimes \iota'_T)$.

10. Using the gasoline demand data of Baltagi and Griffin (1983), given on the Springer web site as GASOLINE.DAT, reproduce Tables 12.1 through 12.7.

11. For the random one-way error components model given in (12.1) and (12.2), consider the OLS estimator of $\text{var}(u_{it}) = \sigma^2$, which is given by $s^2 = e'e/(n - K')$, where e denotes the vector of OLS residuals, $n = NT$ and $K' = K + 1$.

    (a) Show that $E(s^2) = \sigma^2 + \sigma^2_\mu[K' - \text{tr}(I_N \otimes J_T)P_X]/(n - K')$.

    (b) Consider the inequalities given by Kiviet and Krämer (1992) which state that $0 \leq$ mean of $n - K'$ smallest roots of $\Omega \leq E(s^2) \leq$ mean of $n - K'$ largest roots of $\Omega \leq \text{tr}(\Omega)/(n - K')$ where $\Omega = E(uu')$. Show that for the one-way error components model, these bounds are

$$0 \leq \sigma^2_\nu + (n - TK')\sigma^2_\mu/(n - K') \leq E(s^2) \leq \sigma^2_\nu + n\sigma^2_\mu/(n - K') \leq n\sigma^2/(n - K').$$

    As $n \to \infty$, both bounds tend to σ^2, and s^2 is asymptotically unbiased, irrespective of the particular evolution of X, see Baltagi and Krämer (1994) for a proof of this result.

12. Verify the relationship between M and M^*, i.e., $MM^* = M^*$, given below (12.47). **Hint:** Use the fact that $Z = Z^*I^*$ with $I^* = (\iota_N \otimes I_{K'})$.

13. Verify that $\dot{M}$ and $\dot{M}^*$ defined below (12.50) are both symmetric, idempotent and satisfy $\dot{M}\dot{M}^* = \dot{M}^*$.

14. For the gasoline data used in problem 10, verify the Chow-test results given below equation (12.51).

15. For the gasoline data, compute the Breusch-Pagan, Honda and Standardized LM tests for H_0; $\sigma^2_\mu = 0$.

16. If $\widetilde{\beta}$ denotes the LSDV estimator and $\widehat{\beta}_{GLS}$ denotes the GLS estimator, then

    (a) Show that $\widehat{q} = \widehat{\beta}_{GLS} - \widetilde{\beta}$ satisfies $\text{cov}(\widehat{q}, \widehat{\beta}_{GLS}) = 0$.

    (b) Verify equation (12.56).

17. For the gasoline data used in problem 10, replicate the Hausman test results given below equation (12.58).

18. For the cigarette data given as CIGAR.TXT on the Springer web site, reproduce the results given in Table 12.8. See also Baltagi, Griffin and Xiong (2000).

19. *Heteroskedastic Fixed Effects Models.* This is based on Baltagi (1996). Consider the fixed effects model

$$y_{it} = \alpha_i + u_{it} \quad i = 1, 2, \ldots, N; \; t = 1, 2, \ldots, T_i$$

where y_{it} denotes output in industry i at time t and α_i denotes the industry fixed effect. The disturbances u_{it} are assumed to be independent with heteroskedastic variances σ_i^2. Note that the data are unbalanced with different number of observations for each industry.

(a) Show that OLS and GLS estimates of α_i are identical.

(b) Let $\sigma^2 = \sum_{i=1}^{N} T_i \sigma_i^2 / n$ where $n = \sum_{i=1}^{N} T_i$, be the average disturbance variance. Show that the GLS estimator of σ^2 is unbiased, whereas the OLS estimator of σ^2 is biased. Also show that this bias disappears if the data are balanced or the variances are homoskedastic.

(c) Define $\lambda_i^2 = \sigma_i^2/\sigma^2$ for $i = 1, 2 \ldots, N$. Show that for $\alpha' = (\alpha_1, \alpha_2, \ldots, \alpha_N)$

$$\begin{aligned} & E[\text{estimated var}(\widehat{\alpha}_{OLS}) - \text{ true var}(\widehat{\alpha}_{OLS})] \\ = \; & \sigma^2[(n - \sum_{i=1}^{N} \lambda_i^2)/(n-N)] \text{ diag } (1/T_i) - \sigma^2 \text{ diag } (\lambda_i^2/T_i) \end{aligned}$$

This problem shows that in case there are no regressors in the unbalanced panel data model, fixed effects with heteroskedastic disturbances can be estimated by OLS, but one has to correct the standard errors.

20. *The Relative Efficiency of the Between Estimator with Respect to the Within Estimator.* This is based on Baltagi (1999). Consider the simple panel data regression model

$$y_{it} = \alpha + \beta x_{it} + u_{it} \quad i = 1, 2, \ldots, N; \quad t = 1, 2, \ldots, T \tag{1}$$

where α and β are scalars. Subtract the mean equation to get rid of the constant

$$y_{it} - \bar{y}_{..} = \beta(x_{it} - \bar{x}_{..}) + u_{it} - \bar{u}_{..}, \tag{2}$$

where $\bar{x}_{..} = \Sigma_{i=1}^{N} \Sigma_{t=1}^{T} x_{it}/NT$ and $\bar{y}_{..}$ and $\bar{u}_{..}$ are similarly defined. Add and subtract $\bar{x}_{i.}$ from the regressor in parentheses and rearrange

$$y_{it} - \bar{y}_{..} = \beta(x_{it} - \bar{x}_{i.}) + \beta(\bar{x}_{i.} - \bar{x}_{..}) + u_{it} - \bar{u}_{..} \tag{3}$$

where $\bar{x}_{i.} = \Sigma_{t=1}^{T} x_{it}/T$. Now run the unrestricted least squares regression

$$y_{it} - \bar{y}_{..} = \beta_w(x_{it} - \bar{x}_{i.}) + \beta_b(\bar{x}_{i.} - \bar{x}_{..}) + u_{it} - \bar{u}_{..} \tag{4}$$

where β_w is not necessarily equal to β_b.

(a) Show that the least squares estimator of β_w from (4) is the Within estimator and that of β_b is the Between estimator.

(b) Show that if $u_{it} = \mu_i + \nu_{it}$ where $\mu_i \sim \text{IID}(0, \sigma^2_\mu)$ and $\nu_{it} \sim \text{IID}(0, \sigma^2_\nu)$ independent of each other and among themselves, then ordinary least squares (OLS) is equivalent to generalized least squares (GLS) on (4).

(c) Show that for model (1), the relative efficiency of the Between estimator with respect to the Within estimator is equal to $(B_{XX}/W_{XX})[(1-\rho)/(T\rho + (1-\rho))]$, where $W_{XX} = \Sigma^N_{i=1}\Sigma^T_{t=1}(x_{it} - \bar{x}_{i.})^2$ denotes the Within variation and $B_{XX} = T\Sigma^N_{i=1}(\bar{x}_{i.} - \bar{x}_{..})^2$ denotes the Between variation. Also, $\rho = \sigma^2_\mu/(\sigma^2_\mu + \sigma^2_\nu)$ denotes the equicorrelation coefficient.

(d) Show that the square of the t-statistic used to test H_0; $\beta_w = \beta_b$ in (4) yields exactly Hausman's (1978) specification test.

21. For the crime example of Cornwell and Trumbull (1994) studied in Chapter 11. Use the panel data given as CRIME.DAT on the Springer web site to replicate the Between and Within estimates given in Table 1 of Cornwell and Trumbull (1994). Compute 2SLS and Within-2SLS (2SLS with countu dummies) using offense mix and per capita tax revenue as instruments for the probability of arrest and police per capita. Comment on the results.

22. Consider the Arellano and Bond (1991) dynamic employment equation for 140 UK companies over the period 1979-1984. Replicate all the estimation results in Table 4 of Arellano and Bond (1991, p. 290).

References

This chapter is based on Baltagi (2005).

Ahn, S.C. and P. Schmidt (1995), "Efficient Estimation of Models for Dynamic Panel Data," *Journal of Econometrics*, 68: 5-27.

Amemiya, T. (1971), "The Estimation of the Variances in a Variance-Components Model," *International Economic Review*, 12: 1-13.

Anderson, T.W. and C. Hsiao (1982), "Formulation and Estimation of Dynamic Models Using Panel Data, *Journal of Econometrics*, 18: 47-82.

Arellano, M. (1989), "A Note on the Anderson-Hsiao Estimator for Panel Data," *Economics Letters*, 31: 337-341.

Arellano, M. (1993), "On the Testing of Correlated Effects With Panel Data," *Journal of Econometrics*, 59: 87-97.

Arellano, M. and S. Bond (1991), "Some Tests of Specification for Panel Data: Monte Carlo Evidence and An Application to Employment Equations," *Review of Economic Studies*, 58: 277-297.

Balestra, P. (1973), "Best Quadratic Unbiased Estimators of the Variance-Covariance Matrix in Normal Regression," *Journal of Econometrics*, 2: 17-28.

Baltagi, B.H. (1981), "Pooling: An Experimental Study of Alternative Testing and Estimation Procedures in a Two-Way Errors Components Model," *Journal of Econometrics*, 17: 21-49.

Baltagi, B.H. (1996), "Heteroskedastic Fixed Effects Models," Problem 96.5.1, *Econometric Theory*, 12: 867.

Baltagi, B.H. (1999), "The Relative Efficiency of the Between Estimator with Respect to the Within Estimator," Problem 99.4.3, *Econometric Theory*, 15: 630-631.

Baltagi, B.H. (2005), *Econometric Analysis of Panel Data* (Wiley: Chichester).

Baltagi, B.H. and J.M. Griffin (1983), "Gasoline Demand in the OECD: An Application of Pooling and Testing Procedures," *European Economic Review*, 22: 117-137.

Baltagi, B.H., J.M. Griffin and W. Xiong (2000), "To Pool or Not to Pool: Homogeneous Versus Heterogeneous Estimators Applied to Cigarette Demand," *Review of Economics and Statistics*, 82: 117-126.

Baltagi, B.H. and W. Krämer (1994), "Consistency, Asymptotic Unbiasedness and Bounds on the Bias of s^2 in the Linear Regression Model with Error Components Disturbances," *Statistical Papers*, 35: 323-328.

Breusch, T.S. (1987), "Maximum Likelihood Estimation of Random Effects Models," *Journal of Econometrics*, 36: 383-389.

Breusch, T.S. and A.R. Pagan (1980), "The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics," *Review of Economic Studies*, 47: 239-253.

Card (1990), "The Impact of the Mariel Boat Lift on the Miami Labor Market," *Industrial and Labor Relations Review*, 43: 245-253.

Chow, G.C. (1960), "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, 28: 591-605.

Cornwell, C. and W.N. Trumbull (1994), "Estimating the Economic Model of Crime with Panel Data," *Review of Economics and Statistics* 76: 360-366.

Evans, M.A. and M.L. King (1985), "Critical Value Approximations for Tests of Linear Regression Disturbances," *Australian Journal of Statistics*, 27: 68-83.

Fisher, F.M. (1970), "Tests of Equality Between Sets of Coefficients in Two Linear Regressions: An Expository Note," *Econometrica*, 38: 361-366.

Fuller, W.A. and G.E. Battese (1974), "Estimation of Linear Models with Cross-Error Structure," *Journal of Econometrics*, 2: 67-78.

Goldberger, A.S. (1962), "Best Linear Unbiased Prediction in the Generalized Linear Regression Model," *Journal of the American Statistical Association*, 57: 369-375.

Graybill, F.A. (1961), *An Introduction to Linear Statistical Models* (McGraw-Hill: New York).

Hansen, L.P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50: 1029-1054.

Hausman, J.A. (1978), "Specification Tests in Econometrics," *Econometrica*, 46: 1251-1271.

Honda, Y. (1985), "Testing the Error Components Model with Non-Normal Disturbances," *Review of Economic Studies*, 52: 681-690.

Hsiao, C. (2003), *Analysis of Panel Data* (Cambridge University Press: Cambridge).

Judge, G.G., W.E. Griffiths, R.C. Hill, H. Lutkepohl and T.C. Lee (1985), *The Theory and Practice of Econometrics* (Wiley: New York).

Kiviet, J.F. and W. Krämer (1992), "Bias of s^2 in the Linear Regression Model with Correlated Errors," *Empirical Economics*, 16: 375-377.

Maddala, G.S. (1971), "The Use of Variance Components Models in Pooling Cross Section and Time Series Data," *Econometrica*, 39: 341-358.

Maddala, G.S. and T. Mount (1973), "A Comparative Study of Alternative Estimators for Variance Components Models Used in Econometric Applications," *Journal of the American Statistical Association*, 68: 324-328.

Moulton, B.R. and W.C. Randolph (1989), "Alternative Tests of the Error Components Model," *Econometrica*, 57: 685-693.

Nerlove, M. (1971), "A Note on Error Components Models," *Econometrica*, 39: 383-396.

Nickell, S. (1981), "Biases in Dynamic Models with Fixed Effects," *Econometrica*, 49: 1417-1426.

Searle, S.R. (1971), *Linear Models* (Wiley: New York).

Sargan, J. (1958), "The Estimation of Economic Relationships Using Instrumental Variables," *Econometrica*, 26: 393-415.

Swamy, P.A.V.B. and S.S. Arora (1972), "The Exact Finite Sample Properties of the Estimators of Coefficients in the Error Components Regression Models," *Econometrica*, 40: 261-275.

Taub, A.J. (1979), "Prediction in the Context of the Variance-Components Model," *Journal of Econometrics*, 10: 103-108.

Taylor, W.E. (1980), "Small Sample Considerations in Estimation from Panel Data," *Journal of Econometrics*, 13: 203-223.

Wallace, T. and A. Hussain (1969), "The Use of Error Components Models in Combining Cross-Section and Time-Series Data," *Econometrica*, 37: 55-72.

Wansbeek, T.J. and A. Kapteyn (1978), "The Separation of Individual Variation and Systematic Change in the Analysis of Panel Data," *Annales de l'INSEE*, 30-31: 659-680.

Wansbeek, T.J. and A. Kapteyn (1982), "A Simple Way to Obtain the Spectral Decomposition of Variance Components Models for Balanced Data," *Communications in Statistics All*, 2105-2112.

Wansbeek, T.J. and A. Kapteyn, (1989), "Estimation of the error components model with incomplete panels," *Journal of Econometrics* 41: 341–361.

Zellner, A. (1962), "An Efficient Method of Estimating Seemingly Unrelated Regression and Tests for Aggregation Bias," *Journal of the American Statistical Association*, 57: 348-368.

CHAPTER 13
Limited Dependent Variables

13.1 Introduction

In labor economics, one is faced with explaining the decision to participate in the labor force, the decision to join a union, or the decision to migrate from one region to the other. In finance, a consumer defaults on a loan or a credit card debt, or purchases a stock or an asset like a house or a car. In these examples, the dependent variable is usually a dummy variable with values 1 if the worker participates (or consumer defaults on a loan) and 0 if he or she does not participate (or default). We dealt with dummy variables as explanatory variables on the right hand side of the regression, but what additional problems arise when this dummy variable appears on the left hand side of the equation? As we have done in previous chapters, we first study its effects on the usual least squares estimator, and then consider alternative estimators that are more appropriate for models of this nature.

13.2 The Linear Probability Model

What is wrong with running OLS on this model? After all, it is a feasible procedure. For the labor force participation example one regresses the dummy variable for participation on age, sex, race, marital status, number of children, experience and education, etc. The prediction from this OLS regression is interpreted as the likelihood of participating in the labor force. The problems with this interpretation are the following:

(i) We are predicting probabilities of participation for each individual, whereas the actual values observed are 0 or 1.

(ii) There is no guarantee that $\widehat{y}_i$, the predicted value of y_i is going to be between 0 and 1. In fact, one can always find values of the explanatory variables that would generate a corresponding prediction outside the $(0, 1)$ range.

(iii) Even if one is willing to assume that the true model is a linear regression given by

$$y_i = x_i'\beta + u_i \qquad i = 1, 2, \ldots, n. \tag{13.1}$$

what properties does this entail on the disturbances? It is obvious that $y_i = 1$ only when $u_i = 1 - x_i'\beta$, let us say with probability π_i, where π_i is to be determined. Then $y_i = 0$ only when $u_i = -x_i'\beta$ with probability $(1 - \pi_i)$. For the disturbances to have zero mean

$$E(u_i) = \pi_i(1 - x_i'\beta) + (1 - \pi_i)(-x_i'\beta) = 0 \tag{13.2}$$

Solving for π_i, one gets that $\pi_i = x_i'\beta$. This also means that

$$\text{var}(u_i) = \pi_i(1 - \pi_i) = x_i'\beta(1 - x_i'\beta) \tag{13.3}$$

which is heteroskedastic. Goldberger (1964) suggests correcting for this heteroskedasticity by first running OLS to estimate β, and estimating $\sigma_i^2 = \text{var}(u_i)$ by $\widehat{\sigma}_i^2 = x_i'\widehat{\beta}_{OLS}(1 - x_i'\widehat{\beta}_{OLS}) =$

$\widehat{y}_i(1-\widehat{y}_i)$. In the next step a *Weighted Least Squares* (WLS) procedure is run on (13.1) with the original observations divided by $\widehat{\sigma}_i$. One cannot compute $\widehat{\sigma}_i$ if OLS predicts $\widehat{y}_i$ larger than 1 or smaller than 0. Suggestions in the literature include substituting 0.005 instead of $\widehat{y}_i < 0$, and 0.995 for $\widehat{y}_i > 1$. However, these procedures do not perform well, and the WLS predictions themselves are not guaranteed to fall in the $(0, 1)$ range. Therefore, one should use the robust White heteroskedastic variance-covariance matrix option when estimating linear probability models, otherwise the standard errors are biased and inference is misleading.

This brings us to the fundamental problem with OLS, i.e., its functional form. We are trying to predict

$$y_i = F(x_i'\beta) + u_i \tag{13.4}$$

with a linear regression equation, see Figure 13.1, where the more reasonable functional form for this probability is an S-shaped cumulative distribution functional form. This was justified in the biometrics literature as follows: An insect has a tolerance to an insecticide I_i^*, which is an unobserved random variable with *cumulative distribution function* (c.d.f.) F. If the dosage of insecticide administered induces a stimulus I_i that exceeds I_i^*, the insect dies, i.e., $y_i = 1$. Therefore

$$\Pr(y_i = 1) = \Pr(I_i^* \leq I_i) = F(I_i) \tag{13.5}$$

To put it in an economic context, I_i^* could be the unobserved reservation wage of a worker, and if we increase the offered wage beyond that reservation wage, the worker participates in the labor force. In general, I_i could be represented as a function of the individuals characteristics, i.e., the x_i's. $F(x_i'\beta)$ is by definition between zero and 1 for all values of x_i. Also, the linear probability model yields the result that $\partial\pi_i/\partial x_k = \beta_k$, for every i. This means that the probability of participating (π_i) always changes at the same rate with respect to unit increases in the offer wage x_k. However, this probability model gives

$$\partial\pi_i/\partial x_k = [\partial F(z_i)/\partial z_i] \cdot [\partial z_i/\partial x_k] = f(x_i'\beta) \cdot \beta_k \tag{13.6}$$

where $z_i = x_i'\beta$, and f is the *probability density function* (p.d.f.). Equation (13.6) makes more sense because if x_k denotes the offered wage, changing the probability of participation π_i from 0.96 to 0.97 requires a larger change in x_k than changing π_i from 0.23 to 0.24.

If $F(x_i'\beta)$ is the true probability function, assuming it is linear introduces misspecification, and as Figure 13.1 indicates, for $x_i < x_\ell$, all the u_i's generated by a linear probability approximation are positive. Similarly for all $x_i > x_u$, all the u_i's generated by a linear probability approximation are negative.

13.3 Functional Form: Logit and Probit

Having pointed out the problems with considering the functional form F as linear, we turn to two popular functional forms of F, the *logit* and the *probit*. These two c.d.f.'s differ only in the tails, and the logit resembles the c.d.f. of a t-distribution with 7 degrees of freedom, whereas the probit is the normal c.d.f., or that of a t with ∞ degrees of freedom. Therefore, these two forms will give similar predictions unless there are an extreme number of observations in the tails.

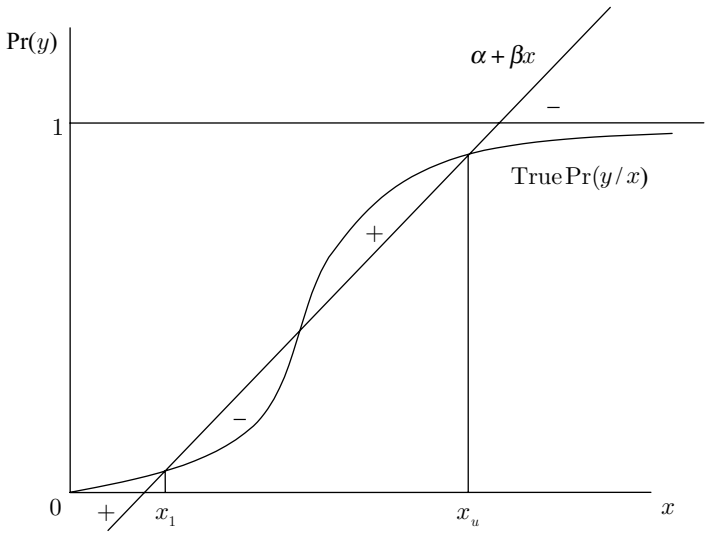


Figure 13.1 Linear Probability Model

We will use the conventional notation $\Phi(z) = \int_{-\infty}^{z} \phi(u)du$, where $\phi(z) = e^{-z^2/2}/\sqrt{2\pi}$ for $-\infty < z < \infty$, for the *probit*. Also, $\Lambda(z) = e^z/(1+e^z) = 1/(1+e^{-z})$ for $-\infty < z < +\infty$, for the *logit*. Some results that we will use quite often in our derivations are the following: $d\Phi/dz = \phi$, and $d\Lambda/dz = \Lambda(1-\Lambda)$. The p.d.f. of the logistic distribution is the product of its c.d.f. and one minus this c.d.f. Therefore, the marginal effects considered above for a general F are respectively,

$$\partial\Phi(x_i'\beta)/\partial x_k = \phi_i\beta_k \tag{13.7}$$

and

$$\partial\Lambda(x_i'\beta)/\partial x_k = \Lambda_i(1-\Lambda_i)\beta_k \tag{13.8}$$

where $\phi_i = \phi(x_i'\beta)$ and $\Lambda_i = \Lambda(x_i'\beta)$.

One has to be careful with the computation of partial derivatives in case there is a dummy variable among the explanatory variables. For such models, one should compute the marginal effects of a change in one unit of a continuous variable x_k for both values of the dummy variable.

Illustrative Example: Using the *probit* model, suppose that the probability of joining a union is estimated as follows: $\widehat{\pi}_i = \Phi(2.5 - 0.06\ WKS_i + 0.95\ OCC_i)$ where WKS is the number of weeks worked and $OCC = 1$, if the individual is in a blue-collar occupation, and zero otherwise. Weeks worked in this sample range from 20 to 50. From (13.7), the marginal effect of one extra week of work on the probability of joining the union is given by:

For Blue-Collar Workers:
$$\begin{aligned} &-0.06\ \phi[2.5 - 0.06\ WKS + 0.95] \\ &= -0.06\ \phi[2.25] = -0.002 \text{ at } WKS = 20 \\ &= -0.06\ \phi[1.35] = -0.010 \text{ at } WKS = 35 \\ &= -0.06\ \phi[0.45] = -0.022 \text{ at } WKS = 50 \end{aligned}$$

For Non Blue-Collar Workers:
$$\begin{aligned} &-0.06\ \phi[2.5 - 0.06\ WKS] \\ &= -0.06\ \phi[1.3] = -0.010 \text{ at } WKS = 20 \\ &= -0.06\ \phi[0.4] = -0.022 \text{ at } WKS = 35 \\ &= -0.06\ \phi[-0.5] = -0.021 \text{ at } WKS = 50 \end{aligned}$$

Note how different these marginal effects are for blue-collar versus non blue-collar workers even for the same weeks worked. Increasing weeks worked from 20 to 21 reduces the probability of joining the union by 0.002 for a Blue-Collar worker. This is compared to five times that amount for a Non Blue-Collar worker.

13.4 Grouped Data

In the biometrics literature, grouped data is very likely from laboratory experiments, see Cox (1970). In the insecticide example, every dosage level x_i is administered to a group of insects of size n_i, and the proportion of insects that die are recorded (p_i). This is done for $i = 1, 2, \ldots, M$ dosage levels.

$$P[y_i = 1] = \pi_i = P[I_i^* \leq I_i] = \Phi(\alpha + \beta x_i)$$

where I_i^* is the tolerance and $I_i = \alpha + \beta x_i$ is the stimulus. In economics, observations may be grouped by income levels or age and we observe the labor participation rate for each income or age group. For this type of grouped data, we estimate the probability of participating in the labor force π_i with p_i, the proportion from the sample. This requires a large number of observations in each group, i.e., a large n_i for $i = 1, 2, \ldots, M$. In this case, the approximation is

$$z_i = \Phi^{-1}(p_i) \cong \alpha + \beta x_i \tag{13.9}$$

for each p_i, we compute the standardized normal variates, the z_i's, and we have an estimate of $\alpha + \beta x_i$. Note that the *standard* normal distribution assumption is not restrictive in the sense that if I_i^* is $N(\mu, \sigma^2)$ rather than $N(0, 1)$, then one standardizes the $P[I_i^* \leq I_i]$ by subtracting μ and dividing by σ, in which case the new I_i^* is $N(0, 1)$ and the new α is $(\alpha - \mu)/\sigma$, whereas the new β is β/σ. This also implies that μ and σ are not separately estimable. A plot of the z_i's versus the x_i's would give estimates of α and β. For the biometrics example, one can compute $LD50$, which is the dosage level that will kill 50% of the insect population. This corresponds to $z_i = 0$, which solves for $x_i = -\widehat{\alpha}/\widehat{\beta}$. Similarly, $LD95$ corresponds to $z_i = 1.645$, which solves for $x_i = (1.645 - \widehat{\alpha})/\widehat{\beta}$. Alternatively, for the economic example, $LD50$ is the minimum reservation wage that is necessary for a 50% labor participation rate.

One could improve on this method by including more x's on the right hand side of (13.9). In this case, one can no longer plot the z_i values versus the x variables. However, one can run

OLS of z on these x's. One problem remains, OLS ignores the heteroskedasticity in the error term. To see this:

$$p_i = \pi_i + \epsilon_i = F(x_i'\beta) + \epsilon_i \tag{13.10}$$

where F is a general c.d.f. and $\pi_i = F(x_i'\beta)$. Using the properties of the binomial distribution, $E(p_i) = \pi_i$ and $\text{var}(p_i) = \pi_i(1-\pi_i)/n_i$. Defining $z_i = F^{-1}(p_i)$, we obtain from (13.10)

$$z_i = F^{-1}(p_i) = F^{-1}(\pi_i + \epsilon_i) \cong F^{-1}(\pi_i) + [dF^{-1}(\pi_i)/d\pi_i]\epsilon_i \tag{13.11}$$

where the approximation $\cong$ is a Taylor series expansion around $\epsilon_i = 0$. Since F is monotonic $\pi_i = F(F^{-1}(\pi_i))$. Let $w_i = F^{-1}(\pi_i) = x_i'\beta$, differentiating with respect to π gives

$$1 = [dF(w_i)/dw_i]dw_i/d\pi_i \tag{13.12}$$

Alternatively, this can be rewritten as

$$dF^{-1}(\pi_i)/d\pi_i = dw_i/d\pi_i = 1/\{dF(w_i)/dw_i\} = 1/f(w_i) = 1/f(x_i'\beta) \tag{13.13}$$

where f is the probability density function corresponding to F. Using (13.13), equation (13.11) can be rewritten as

$$\begin{aligned} z_i &= F^{-1}(p_i) \cong F^{-1}(\pi_i) + \epsilon_i/f(x_i'\beta) \\ &= F^{-1}(F(x_i'\beta)) + \epsilon_i/f(x_i'\beta) = x_i'\beta + \epsilon_i/f(x_i'\beta) \end{aligned} \tag{13.14}$$

From (13.14), it is clear that the regression disturbances of z_i on x_i are given by $u_i \cong \epsilon_i/f(x_i'\beta)$, with $E(u_i) = 0$ and $\sigma_i^2 = \text{var}(u_i) = \text{var}(\epsilon_i)/f^2(x_i'\beta) = \pi_i(1-\pi_i)/(n_i f_i^2) = F_i(1-F_i)/(n_i f_i^2)$ since $\pi_i = F_i$ where the subscript i on f or F denotes that the argument of that function is $x_i'\beta$. This heteroskedasticity in the disturbances renders OLS on (13.14) consistent but inefficient. For the *probit*, $\sigma_i^2 = \Phi_i(1-\Phi_i)/(n_i\phi_i^2)$, and for the *logit*, $\sigma_i^2 = 1/[n_i\Lambda_i(1-\Lambda_i)]$, since $f_i = \Lambda_i(1-\Lambda_i)$. Using $1/\sigma_i$ as weights, a WLS procedure can be performed on (13.14). Note that $F^{-1}(p)$ for the *logit* is simply $\log[p/(1-p)]$. This is one more reason why the logistic functional form is so popular. In this case one regresses $\log[p/(1-p)]$ on x correcting for heteroskedasticity using WLS. This procedure is also known as the minimum logit chi-square method and is due to Berkson (1953).

In order to obtain feasible estimates of the σ_i's, one could use the OLS estimates of β from (13.14), to estimate the weights. Greene (1993) argues that one should *not* use the proportions p_i's as estimates for the π_i's because this is equivalent to using the y_i^2's instead of σ_i^2 in the heteroskedastic regression. These will lead to inefficient estimates. If OLS on (13.14) is reported one should use the robust White heteroskedastic variance-covariance option, otherwise the standard errors are biased and inference is misleading.

Ruhm (1996) uses a grouped logit analysis to study the impact of beer taxes and a variety of alcohol-control policies on motor vehicle fatality rates. Ruhm uses panel data of 48 states (excluding Alaska, Hawaii and the District of Columbia) over the period 1982-1988. The dependent variable is $\log[p/(1-p)]$ where p is the total vehicle fatality rate per capita for state i at time t. The explanatory variables include the real beer tax rate on 24 (12 oz.) containers of beer, the minimum legal drinking age (MLDA) in years and five other dummy variables indicating the presence of alcohol regulations. These include BREATH test laws which is a dummy variable that takes the value 1 if the state authorized the police to administer pre-arrest breath

test to establish probable cause for driving under the influence (DUI). DRAMLAW is dummy variable taking the value 1 if the state has a statute or case law authorizing parties injured by an intoxicated driver to file a law suit against the alcohol server. PER SE takes the value 1 if the state licensing agency is required to suspend or revoke the driver's license after arrest for DUI. CONSENT takes the value 1 if the state has a law requiring license sanction for refusing to submit to alcohol testing. JAIL which takes the value of 1 if the state passed legislation mandating jail or community service for the first DUI conviction. Other variables included are the unemployment rate, real per capita income and state and time dummy variables. This is the *fixed effects* specification discussed in Chapter 12. Details on the definitions of these variables are given in Table 1 of Ruhm (1996). Results showed that most of the regulations had little or no impact on traffic mortality. By contrast, higher beer taxes were associated with reductions in crash deaths.

Table 13.1 shows the grouped logit regression results for the Ruhm (1996) data set using state and time dummy variables. The beer tax is negative and significant, while the minimum legal drinking age is not significant. Neither is the breath test law, the PER SE, CONSENT or JAIL variables, all of which represent state alcohol safety related legislation. However, DRAMLAW is negative and significant. Income per capita is positive and significant while the unemployment rate is negative and significant. Lower unemployment and higher income are signs of good economic conditions with the associated higher traffic density. Also, higher income per capita leads to higher consumption of alcohol, and higher unemployment leads to more depressed and unsatisfied labor force which could encourage more alcohol consumption and traffic related deaths. The state dummy variables are jointly significant with an observed F-value of 48.85 which is distributed as $F(47, 272)$. The year dummies are jointly significant with an observed F-value of 8.01 which is distributed as $F(6, 272)$. Problem 12 asks the reader to replicate Table 13.1. These results imply that increasing the minimum legal drinking age, revoking driver's license or imposing stiffer punishments like mandating jail or community service are not effective policy tools for decreasing traffic related deaths. However, increasing the real tax on beer is an effective policy for reducing traffic related deaths.

For grouped data, the sample sizes n_i for each group have to be sufficiently large. Also, the p_i's cannot be zero or one. One modification suggested in the literature is to add $(1/2n_i)$ to p_i when computing the log of odds ratio, see Cox (1970).

Papke and Wooldridge (1996) argue that in many economic settings p_i may be 0 or 1 for a large number of observations. For example, when studying participation rates in pension plans or when studying high school graduation rates. They propose a fractional logit regression which handles fractional response variables based on quasi-likelihood methods. Fractional response variables are bounded variables. Without loss of generality, they could be restricted to lie between 0 and 1. Examples include the proportion of income spent on charitable contributions, the fraction of total weekly hours spent working. Papke and Wooldridge (1996) propose modeling the $E(y_i/x_i)$ as a logistic function $\Lambda(x_i'\beta)$. This insures that the predicted value of y_i lies in the interval $(0, 1)$. It is also well defined even if y_i takes the values 0 or 1 with positive probability. It is important to note that in case y_i is a proportion from a group of known size n_i, the quasi maximum likelihood method ignores the information on n_i. Using the Bernoulli log-likelihood function, one gets

$$L_i(\beta) = y_i \log[\Lambda(x_i'\beta)] + (1 - y_i) \log[1 - \Lambda(x_i'\beta)]$$

for $i = 1, 2, \ldots, n$, with $0 < \Lambda(x_i'\beta) < 1$.

Table 13.1 Grouped Logit, Beer Tax and Motor Vehicle Fatality

Fixed-effects (within) regression		Number of obs =	335
Group variable (i): state		Number of groups =	48
R-sq within = 0.3566		Obs per group: min =	6
between = 0.2728		avg =	7.0
overall = 0.1735		max =	7
		F(15,272) =	10.05
corr(u_i, Xb) = -0.7400		Prob > F =	0.0000

vfrall	Coef.	Std. Err.	t	$P > t$	[95% Conf.	Interval]
beertax	−.1609269	.0782078	−2.06	0.041	−.3148965	−.0069573
mlda	−.0008891	.0085417	−0.10	0.917	−.0177053	.0159272
breath	.0018155	.0226649	0.08	0.936	−.0428055	.0464365
dramlaw	−.0789037	.0246141	−3.21	0.002	−.1273621	−.0304453
perse	−.0192373	.0199354	−0.96	0.335	−.0584845	.0200099
impcond	.0232511	.0386605	0.60	0.548	−.0528608	.099363
jailcom	.0058019	.0282758	0.21	0.838	−.0498654	.0614693
incperc	.0246669	.0102202	2.41	0.016	.0045461	.0447877
unrate	−.0335797	.0050031	−6.71	0.000	−.0434294	−.0237299
year83	−.0336748	.0156339	−2.15	0.032	−.0644536	−.002896
year84	−.1028847	.0194028	−5.30	0.000	−.1410836	−.0646859
year85	−.1348613	.0205257	−6.57	0.000	−.1752707	−.0944519
year86	−.1005879	.0224115	−4.49	0.000	−.14471	−.0564658
year87	−.1337847	.0258383	−5.18	0.000	−.1846531	−.0829162
year88	−.159083	.0292153	−5.45	0.000	−.2165999	−.1015661
_cons	−8.394118	.2275537	−36.89	0.000	−8.842108	−7.946127
sigma_u	.36371961					
sigma_e	.07263869					
rho	.9616454	(fraction of variance due to u_i)				
F test that all u_i=0:		F(47, 272) = 48.85			Prob > F =	0.0000

Maximizing $\sum_{i=1}^{n} L_i(\beta)$ with respect to β yields the quasi-MLE which is consistent and $\sqrt{n}$ asymptotically normal *regardless* of the distribution of y_i conditional on x_i, see Gourieroux, Monfort and Trognon (1984) and McCullagh and Nelder (1989). The latter proposed the generalized linear models (GLM) approach to this problem in statistics. Logit QMLE can be done in Stata using the GLM command with the Binary family function indicating Bernoulli and the Link function indicating the logistic distribution.

Papke and Wooldridge (1996) derive robust asymptotic variance of the QMLE of β and suggest some specification tests based on Wooldridge (1991). They apply their methods to the participation in 401(K) pension plans. The data are from the 1987 IRS Form 5500 reports of pension plans with more than 100 participants. This data set containing 4734 observations can be downloaded from the *Journal of Applied Econometrics* Data Archive. We focus on a subset of their data which includes 3874 observations of plans with match rates less than or equal to one. Match rates above one may be indicating end-of-plan year employer contributions made to avoid IRS disqualification. Participation rates (PRATE) in this sample are high

Table 13.2 Logit Quasi-MLE of Participation Rates in 401(K) Plan

glm prate mrate log_emp log_emp2 age age2 sole if one==1, f(bin) l(logit) robust
note: prate has non-integer values

Iteration 0: log pseudo-likelihood = −1200.8698
Iteration 1: log pseudo-likelihood = −1179.3843
Iteration 2: log pseudo-likelihood = −1179.2785
Iteration 3: log pseudo-likelihood = −1179.2785

Generalized linear models		Number of obs	= 3784
Optimization	: ML: Newton-Raphson	Residual df	= 3777
		Scale parameter	= 1
Deviance	= 1273.60684	(1/df) Deviance	= .3372006
Pearson	= 724.4199889	(1/df) Pearson	= .1917977
Variance function	: V(u) = u*(1-u)	[Bernoulli]	
Link function	: g(u) = ln(u/(1-u))	[Logit]	
Standard errors	: Sandwich		
Log pseudo-likelihood	= −1179.278516		
BIC	= −29843.34715	AIC	= .6269971

prate	Coef.	Robust Std. Err.	z	$P>\|z\|$	[95% Conf.	Interval]
mrate	1.39008	.1077064	12.91	0.000	1.17898	1.601181
log_emp	−1.001874	.1104365	−9.07	0.000	−1.218326	−.7854229
log_emp2	.0521864	.0071278	7.32	0.000	.0382161	.0661568
age	.0501126	.0088451	5.67	0.000	.0327766	.0674486
age2	−.0005154	.0002117	−2.43	0.015	−.0009303	−.0001004
sole	.0079469	.0502025	0.16	0.874	−.0904482	.1063421
_cons	5.057997	.4208646	12.02	0.000	4.233117	5.882876

averaging 84.8%. Over 40% of the plans have a participation proportion of one. This makes the log-odds ratio approach awkward since adjustments have to be made to more than 40% of the observations. The plan match rate (MRATE) averages about 41 cents on the dollar. Other explanatory variables include total firm employment (EMP), age of the plan (AGE), a dummy variable (SOLE) which takes the value of 1 if the 401(K) plan is the only pension plan offered by the employer. The 401(K) plans average 12 years in age, they are the SOLE plan in 37% of the sample. The average employment is 4622. Problem 14 asks the reader to replicate the descriptive statistic given in Table I of Papke and Wooldridge (1996, p. 627). Table 13.2 gives the Stata output for logit QMLE using the same specification given in Table II of Papke and Wooldridge (1996, p. 628). Note that it uses the GLM command, the Bernoulli variance function and the logit link function. The results show that there is a positive and significant relationship between match rate and participation rate. All the other variables included are significant except for SOLE. Problem 14 asks the reader to replicate this result and compare with OLS. The latter turns out to have a lower R^2 and fails a RESET test, see Chapter 8.

13.5 Individual Data: Probit and Logit

When the number of observations n_i in each group is small, one cannot obtain reliable estimates of the π_i's with the p_i's. In this case, one should not group the observations, instead these observations should be treated as individual observations and the model estimated by the maximum likelihood procedure. The likelihood is obtained as independent random draws from a Bernoulli distribution with probability of success $\pi_i = F(x_i'\beta) = P[y_i = 1]$. Hence

$$\ell = \prod_{i=1}^{n} [F(x_i'\beta)]^{y_i} [1 - F(x_i'\beta)]^{1-y_i} \tag{13.15}$$

and the log-likelihood

$$\log\ell = \sum_{i=1}^{n} \{y_i \log F(x_i'\beta) + (1 - y_i)\log[1 - F(x_i'\beta)]\} \tag{13.16}$$

The first-order conditions for maximization require the score $S(\beta) = \partial\log\ell/\partial\beta$ to be zero:

$$\begin{aligned} S(\beta) &= \partial\log\ell/\partial\beta = \sum_{i=1}^{n} \{[f_i y_i/F_i] - (1 - y_i)[f_i/(1 - F_i)]\} x_i \\ &= \sum_{i=1}^{n} (y_i - F_i) f_i x_i/[F_i(1 - F_i)] = 0 \end{aligned} \tag{13.17}$$

where the subscript i on f or F denotes that the argument of that function is $x_i'\beta$. For the logit model (13.17) reduces to

$$S(\beta) = \sum_{i=1}^{n} (y_i - \Lambda_i) x_i = 0 \qquad \text{since} \qquad f_i = \Lambda_i(1 - \Lambda_i) \tag{13.18}$$

If there is a constant in the model, the solution to (13.18) for $x_i = 1$ implies that $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{\Lambda}_i$. This means that the number of participants in the sample, i.e., those with $y_i = 1$, will always be equal to the predicted number of participants from the logit model. Similarly, if x_i is a dummy variable which is 1 if the individual is male and zero if the individual is female, then (13.18) states that the predicted frequency is equal to the actual frequency for males and females. Note that (13.18) resembles the OLS normal equations if we interpret $(y_i - \widehat{\Lambda}_i)$ as residuals. For the probit model (13.17) reduces to

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n} (y_i - \Phi_i)\phi_i x_i/[\Phi_i(1 - \Phi_i)] \\ &= \sum_{y_i=0} \lambda_{oi} x_i + \sum_{y_i=1} \lambda_{1i} x_i = 0 \end{aligned} \tag{13.19}$$

where $\lambda_{oi} = -\phi_i/[1-\Phi_i]$ for $y_i = 0$ and $\lambda_{1i} = \phi_i/\Phi_i$ for $y_i = 1$. Also, $\Sigma_{y_i=0}$ denotes the sum over all zero values of y_i. These λ_i's are thought of as *generalized residuals* which are orthogonal to x_i. Note that unlike the logit, the probit does not necessarily predict the number of participants to be exactly equal to the number of ones in the sample.

Equations (13.17) are highly nonlinear and may be solved using the scoring method, i.e., starting with some initial value β_o we revise this estimate as follows:

$$\beta_1 = \beta_o + [I^{-1}(\beta_o)]S(\beta_o) \tag{13.20}$$

where $S(\beta) = \partial\log\ell/\partial\beta$ and $I(\beta) = E[-\partial^2\log\ell/\partial\beta\partial\beta']$. This process is repeated until convergence. For the logit and probit models, $\log F(x_i'\beta)$ and $\log[1 - F(x_i'\beta)]$ are concave. Hence, the log-likelihood function given by (13.16) is *globally concave*, see Pratt (1981). Hence, for both the logit and probit, $[\partial^2\log\ell/\partial\beta\partial\beta']$ is negative definite for all values of β and the iterative procedure will converge to the unique maximum likelihood estimate $\widehat{\beta}_{MLE}$ no matter what

starting values we use. In this case, the asymptotic covariance matrix of $\widehat{\beta}_{MLE}$ is estimated by $I^{-1}(\widehat{\beta}_{MLE})$ from the last iteration.

Amemiya (1981, p. 1495) derived $I(\beta)$ by differentiating (13.17), multiplying by a negative sign and taking the expected value, the result is given by:

$$I(\beta) = -E[\partial^2 \log \ell / \partial\beta\partial\beta'] = \sum_{i=1}^{n} f_i^2 x_i x_i' / F_i(1-F_i) \tag{13.21}$$

For the logit, (13.21) reduces to

$$I(\beta) = \sum_{i=1}^{n} \Lambda_i(1-\Lambda_i) x_i x_i' \tag{13.22}$$

For the probit, (13.21) reduces to

$$I(\beta) = \sum_{i=1}^{n} \phi_i^2 x_i x_i' / \Phi_i(1-\Phi_i) \tag{13.23}$$

Alternative maximization may use the Newton-Raphson iterative procedure which uses the Hessian itself rather than its expected value in (13.20), i.e., $I(\beta)$ is replaced by $H(\beta) = [-\partial^2 \log \ell / \partial\beta\partial\beta']$. For the logit model, $H(\beta) = I(\beta)$ and is given in (13.22). For the probit model, $H(\beta) = \sum_{i=1}^{n} [\lambda_i^2 + \lambda_i x_i'\beta] x_i x_i'$ which is different from (13.23). Note that $\lambda_i = \lambda_{oi}$ if $y_i = 0$; and $\lambda_i = \lambda_{1i}$ if $y_i = 1$. These were defined below (13.19).

A third method, suggested by Berndt, Hall, Hall and Hausman (1974) uses the outer product of the first derivatives in place of $I(\beta)$, i.e., $G(\beta) = S(\beta)S'(\beta)$. For the logit model, this is $G(\beta) = \sum_{i=1}^{n} (y_i - \Lambda_i)^2 x_i x_i'$. For the probit model, $G(\beta) = \sum_{i=1}^{n} \lambda_i^2 x_i x_i'$. As in the method of scoring, one iterates starting from initial estimates β_o, and the asymptotic variance-covariance matrix is estimated from the inverse of $G(\widehat{\beta})$, $H(\widehat{\beta})$ or $I(\widehat{\beta})$ in the last iteration.

Test of hypotheses can be carried out from the asymptotic standard errors using t-statistics. For $R\beta = r$ type restrictions, the usual Wald test $W = (R\widehat{\beta} - r)'[RV(\widehat{\beta})R']^{-1}(R\widehat{\beta} - r)$ can be used with $V(\widehat{\beta})$ obtained from the last iteration as described above. Likelihood ratio and Lagrange Multiplier statistics can also be computed. $LR = -2[\log \ell_{restricted} - \log \ell_{unrestricted}]$, whereas, the Lagrange Multiplier statistic is $LM = S'(\beta)V(\beta)S(\beta)$, where $S(\beta)$ is the score evaluated at the restricted estimator. Davidson and MacKinnon (1984) suggest that $V(\widehat{\beta})$ based on $I(\widehat{\beta})$ is the best of the three estimators to use. In fact, Monte Carlo experiments show that the estimate of $V(\widehat{\beta})$ based on the outer product of the first derivatives usually performs the worst and is not recommended in practice. All three statistics are asymptotically equivalent and are asymptotically distributed as χ_q^2 where q is the number of restrictions. The next section discusses tests of hypotheses using an artificial regression.

13.6 The Binary Response Model Regression[1]

Davidson and MacKinnon (1984) suggest a modified version of the Gauss-Newton regression (GNR) considered in Chapter 8 which is useful in the context of a binary response model described in (13.5).[2] In fact, we have shown that this model can be written as a nonlinear regression

$$y_i = F(x_i'\beta) + u_i \tag{13.24}$$

with u_i having zero mean and $\text{var}(u_i) = F_i(1-F_i)$. The GNR ignoring heteroskedasticity yields

$$(y_i - F_i) = f_i x_i' b + \text{residual}$$

where b is the regression estimates when we regress $(y_i - F_i)$ on $f_i x_i'$.

Correcting for heteroskedasticity by dividing each observation by its standard deviation we get the *Binary Response Model Regression* (BRMR):

$$\frac{(y_i - F_i)}{\sqrt{F_i(1-F_i)}} = \frac{f_i}{\sqrt{F_i(1-F_i)}}\, x_i' b + \text{residual} \tag{13.25}$$

For the logit model with $f_i = \Lambda_i(1-\Lambda_i)$, this simplifies further to

$$\frac{y_i - \Lambda_i}{\sqrt{f_i}} = \sqrt{f_i}\, x_i' b + \text{residual} \tag{13.26}$$

For the probit model, the BRMR is given by

$$\frac{y_i - \Phi_i}{\sqrt{\Phi_i(1-\Phi_i)}} = \frac{\phi_i}{\sqrt{\Phi_i(1-\Phi_i)}}\, x_i' b + \text{residual} \tag{13.27}$$

Like the GNR considered in Chapter 8, the BRMR given in (13.25) can be used for obtaining parameter and covariance matrix estimates as well as test of hypotheses. In fact, Davidson and MacKinnon point out that the transpose of the dependent variable in (13.25) times the matrix of regressors in (13.25) yields a vector whose typical element is exactly that of $S(\beta)$ given in (13.17). Also, the transpose of the matrix of regressors in (13.25) multiplied by itself yields a matrix whose typical element is exactly that of $I(\beta)$ given in (13.21).

Let us consider how the BRMR is used to test hypotheses. Suppose that $\beta' = (\beta_1', \beta_2')$ where β_1 is of dimension $k-r$ and β_2 is of dimension r. We want to test H_o; $\beta_2 = 0$. Let $\widetilde{\beta}' = (\widetilde{\beta}_1, 0)$ be the restricted MLE of β subject to H_o. In order to test H_o, we run the BRMR:

$$\frac{y_i - \widetilde{F}_i}{\sqrt{\widetilde{F}_i(1-\widetilde{F}_i)}} = \frac{\widetilde{f}_i}{\sqrt{\widetilde{F}_i(1-\widetilde{F}_i)}}\, x_{i1}' b_1 + \frac{\widetilde{f}_i}{\sqrt{\widetilde{F}_i(1-\widetilde{F}_i)}}\, x_{i2}' b_2 + \text{residual} \tag{13.28}$$

where $x_i' = (x_{i1}', x_{i2}')$ has been partitioned into vectors conformable with the corresponding partition of β. Also, $\widetilde{F}_i = F(x_i'\widetilde{\beta})$ and $\widetilde{f}_i = f(x_i'\widetilde{\beta})$. The suggested test statistic for H_o is the explained sum of squares of the regression (13.28). This is asymptotically distributed as χ_r^2 under H_o.[3] A special case of this BRMR is that of testing the null hypothesis that *all* the slope coefficients are zero. In this case, $x_{i1} = 1$ and β_1 is the constant α. Problem 2 shows that the restricted MLE in this case is $\widetilde{F}(\alpha) = \bar{y}$ or $\widetilde{\alpha} = F^{-1}(\bar{y})$, where $\bar{y}$ is the proportion of the sample with $y_i = 1$. Therefore, the BRMR in (13.25) reduces to

$$\frac{y_i - \bar{y}}{\sqrt{\bar{y}(1-\bar{y})}} = \frac{f_i(\widetilde{\alpha})}{\sqrt{\bar{y}(1-\bar{y})}}\, b_1 + \frac{f_i(\widetilde{\alpha})}{\sqrt{\bar{y}(1-\bar{y})}}\, x_{i2}' b_2 + \text{residual} \tag{13.29}$$

Note that $\bar{y}(1-\bar{y})$ is *constant* for all observations. The test for $b_2 = 0$ is not affected by dividing the dependent variable or the regressors by this constant, nor is it affected by subtracting a constant from the dependent variable. Hence, the test for $b_2 = 0$ can be carried out by regressing y_i on a constant and x_{i2} and testing that the slope coefficients of x_{i2} are zero using the usual least squares F-statistic. This is a simpler alternative to the likelihood ratio test proposed in the previous section and described in the empirical example in section 13.9. For other uses of the BRMR, see Davidson and MacKinnon (1993).

13.7 Asymptotic Variances for Predictions and Marginal Effects

Two results of interest after estimating the model are: the *predictions* $F(x'\widehat{\beta})$ and the *marginal effects* $\partial F/\partial x = f(x'\widehat{\beta})\widehat{\beta}$. For example, given the characteristics of an individual x, we can predict his or her probability of purchasing a car. Also, given a change in x, say income, one can estimate the marginal effect this will have on the probability of purchasing a car. The latter effect is constant for the linear probability model and is given by the regression coefficient of income, whereas for the probit and logit models this marginal effect will vary with the x_i's, see (13.7) and (13.8). These marginal effects can be computed with Stata using the *dprobit* command. The default is to compute them at the sample mean $\overline{x}$. There is also the additional problem of computing variances for these predictions and marginal effects. Both $F(x'\widehat{\beta})$ and $f(x'\widehat{\beta})\widehat{\beta}$ are nonlinear functions of the $\widehat{\beta}$'s. To compute standard errors, we can use the following linear approximation which states that whenever $\widehat{\theta} = F(\widehat{\beta})$ then the asy.var$(\widehat{\theta}) = (\partial F/\partial\widehat{\beta})'V(\widehat{\beta})(\partial F/\partial\widehat{\beta})$. For the predictions, let $z = x'\widehat{\beta}$ and denote by $F = F(x'\widehat{\beta})$ and $\widehat{f} = f(x'\widehat{\beta})$, then

$$\partial\widehat{F}/\partial\widehat{\beta} = (\partial\widehat{F}/\partial z)(\partial z/\partial\widehat{\beta}) = \widehat{f}x \qquad \text{and} \qquad \text{asy.var}(\widehat{F}) = \widehat{f}^2 x'V(\widehat{\beta})x.$$

For the marginal effects, let $\widehat{\gamma} = \widehat{f}\widehat{\beta}$, then

$$\text{asy.var}(\widehat{\gamma}) = (\partial\widehat{\gamma}/\partial\widehat{\beta}')V(\widehat{\beta})(\partial\widehat{\gamma}/\partial\widehat{\beta}')' \tag{13.30}$$

where $\partial\widehat{\gamma}/\partial\widehat{\beta}' = \widehat{f}I_k + \widehat{\beta}(\partial\widehat{f}/\partial z)(\partial z/\partial\widehat{\beta}') = \widehat{f}I_k + (\partial\widehat{f}/\partial z)(\widehat{\beta}x')$.

For the probit model, $\partial\widehat{f}/\partial z = \partial\widehat{\phi}/\partial z = -z\widehat{\phi}$. So, $\partial\widehat{\gamma}/\partial\widehat{\beta}' = \widehat{\phi}[I_k - z\widehat{\beta}x']$ and

$$\text{asy.var}(\widehat{\gamma}) = \widehat{\phi}^2[I_k - x'\widehat{\beta}\widehat{\beta}x']V(\widehat{\beta})[I_k - x'\widehat{\beta}\widehat{\beta}x']' \tag{13.31}$$

For the logit model, $\widehat{f} = \widehat{\Lambda}(1-\widehat{\Lambda})$, so

$$\begin{aligned} \partial\widehat{f}/\partial z &= (1-2\widehat{\Lambda})(\partial\widehat{\Lambda}/\partial z) = (1-2\widehat{\Lambda})(\widehat{f}) = (1-2\widehat{\Lambda})\widehat{\Lambda}(1-\widehat{\Lambda}) \\ \partial\widehat{\gamma}/\partial\widehat{\beta}' &= \widehat{\Lambda}(1-\widehat{\Lambda})[I_k + (1-2\widehat{\Lambda})\widehat{\beta}x'] \end{aligned}$$

and (13.30) becomes

$$\text{asy.var}(\widehat{\gamma}) = [\widehat{\Lambda}(1-\widehat{\Lambda})]^2[I_k + (1-2\widehat{\Lambda})\widehat{\beta}x']V(\widehat{\beta})[I_k + (1-2\widehat{\Lambda})\widehat{\beta}x']' \tag{13.32}$$

13.8 Goodness of Fit Measures

There are problems with the use of conventional R^2-type measures when the explained variable y takes on two values, see Maddala (1983, pp. 37-41). The predicted values $\widehat{y}$ are probabilities and the actual values of y are either 0 or 1 so the usual R^2 is likely to be very low. Also, if there is a constant in the model the linear probability and logit models satisfy $\sum_{i=1}^n y_i = \sum_{i=1}^n \widehat{y}_i$. However, the probit model does not necessarily satisfy this exact relationship although it is approximately valid.

Several R^2-type measures have been suggested in the literature, some of these are the following:

(i) The squared correlation between y and $\widehat{y}$: $R_1^2 = r_{y,\widehat{y}}^2$.

(ii) Measures based on the *residual sum of squares*: Effron (1978) suggested using

$$R_2^2 = 1 - [\sum_{i=1}^n (y_i - \widehat{y}_i)^2 / \sum_{i=1}^n (y_i - \bar{y})^2] = 1 - [n \sum_{i=1}^n (y_i - \widehat{y}_i)^2 / n_1 n_2]$$

since $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2 = n_1 - n(n_1/n)^2 = n_1 n_2/n$, where $n_1 = \sum_{i=1}^n y_i$ and $n_2 = n - n_1$.

Amemiya (1981, p. 1504) suggests using $\left[\sum_{i=1}^n (y_i - \widehat{y}_i)^2 / \widehat{y}_i(1 - \widehat{y}_i)\right]$ as the residual sum of squares. This weights each squared error by the inverse of its variance.

(iii) Measures based on likelihood ratios: $R_3^2 = 1 - (\ell_r/\ell_u)^{2/n}$ where ℓ_r is the restricted likelihood and ℓ_u is the unrestricted likelihood. This tests that all the slope coefficients are zero in the standard linear regression model. For the limited dependent variable model however, the likelihood function has a maximum of 1. This means that $\ell_r \leq \ell_u \leq 1$ or $\ell_r \leq (\ell_r/\ell_u) \leq 1$ or $\ell_r^{2/n} \leq 1 - R_3^2 \leq 1$ or $0 \leq R_3^2 \leq 1 - \ell_r^{2/n}$. Hence, Cragg and Uhler (1970) suggest a pseudo-R^2 that lies between 0 and 1, and is given by $R_4^2 = (\ell_u^{2/n} - \ell_r^{2/n})/[(1 - \ell_r^{2/n})/\ell_u^{2/n}]$. Another measure suggested by McFadden (1974) is $R_5^2 = 1 - (\log\ell_u/\log\ell_r)$.

(iv) Proportion of correct predictions: After computing $\widehat{y}$, one classifies the i-th observation as a success if $\widehat{y}_i > 0.5$, and a failure if $\widehat{y}_i < 0.5$. This measure is useful but may not have enough discriminatory power.

13.9 Empirical Examples

Example 1: Union Participation

To illustrate the logit and probit models, we consider the PSID data for 1982 used in Chapter 4. In this example, we are interested in modelling union participation. Out of the 595 individuals observed in 1982, 218 individuals had their wage set by a union and 377 did not. The explanatory variables used are: years of education (ED), weeks worked (WKS), years of full-time work experience (EXP), occupation ($OCC = 1$, if the individual is in a blue-collar occupation), residence ($SOUTH = 1$, $SMSA = 1$, if the individual resides in the South, or in a standard metropolitan statistical area), industry ($IND = 1$, if the individual works in a manufacturing industry), marital status ($MS = 1$, if the individual is married), sex and race ($FEM = 1$, $BLK = 1$, if the individual is female or black). A full description of the data is given in Cornwell and Rupert (1988). The results of the linear probability, logit and probit models are given in Table 13.3. These were computed using EViews. In fact Table 13.4 gives the probit output. We have already mentioned that the probit model normalizes σ to be 1. But, the logit model has variance $\pi^2/3$. Therefore, the logit estimates tend to be larger than the probit estimates although by a factor less than $\pi/\sqrt{3}$. In order to make the logit results comparable to those of the probit, Amemiya (1981) suggests multiplying the logit coefficient estimates by 0.625.

Similarly, to make the linear probability estimates comparable to those of the probit model one needs to multiply these coefficients by 2.5 and then subtract 1.25 from the constant term. For this example, both logit and probit procedures converged quickly in 4 iterations. The log-likelihood values and McFadden's (1974) R^2 obtained for the last iteration are recorded.

Note that the logit and probit estimates yield similar results in magnitude, sign and significance. One would expect different results from the logit and probit only if there are several observations in the tails. The following variables were insignificant at the 5% level: EXP, IND,

Table 13.3 Comparison of the Linear Probability, Logit and Probit Models: Union Participation*

Variable	OLS	Logit	Probit
EXP	−.005 (1.14)	−.007 (1.15)	−.007 (1.21)
WKS	−.045 (5.21)	−.068 (5.05)	−.061 (5.16)
OCC	.795 (6.85)	1.036 (6.27)	.955 (6.28)
IND	.075 (0.79)	.114 (0.89)	.093 (0.76)
SOUTH	−.425 (4.27)	−.653 (4.33)	−.593 (4.26)
SMSA	.211 (2.20)	.280 (2.05)	.261 (2.03)
MS	.247 (1.55)	.378 (1.66)	.351 (1.62)
FEM	−.272 (1.37)	−.483 (1.58)	−.407 (1.47)
ED	−.040 (1.88)	−.057 (1.85)	−.057 (1.99)
BLK	.125 (0.71)	.222 (0.90)	.226 (0.99)
Const	1.740 (5.27)	2.738 (3.27)	2.517 (3.30)
Log-likelihood		−312.337	−313.380
McFadden's R^2		0.201	0.198
χ^2_{10}		157.2	155.1

* Figures in parentheses are t-statistics

MS, FEM and BLK. The results show that union participation is less likely if the individual resides in the South and more likely if he or she resides in a standard metropolitan statistical area. Union participation is also less likely the more the weeks worked and the higher the years of education. Union participation is more likely for blue-collar than non blue-collar occupations. The linear probability model yields different estimates from the logit and probit results. OLS predicts two observations with $\widehat{y}_i > 1$, and 29 observations with $\widehat{y}_i < 0$. Table 13.5 gives the actual versus predicted values of union participation for the linear probability, logit and probit models. The percentage of correct predictions is 75% for the linear probability and probit model and 76% for the logit model.

One can test the significance of all slope coefficients by computing the LR based on the unrestricted log-likelihood value ($\log\ell_u$) reported in Table 13.3, and the restricted log-likelihood value including only the constant. The latter is the same for both the logit and probit models and is given by

$$\log\ell_r = n[\bar{y}\log\bar{y} + (1-\bar{y})\log(1-\bar{y})] \tag{13.33}$$

where $\bar{y}$ is the proportion of the sample with $y_i = 1$, see problem 2. In this example, $\bar{y} = 218/595 = 0.366$ and $n = 595$ with $\log\ell_r = -390.918$. Therefore, for the probit model,

$$LR = -2[\log\ell_r - \log\ell_u] = -2[-390.918 + 313.380] = 155.1$$

which is distributed as χ^2_{10} under the null of zero slope coefficients. This is highly significant and the null is rejected. Similarly, for the logit model this LR statistic is 157.2. For the linear probability model, the same null hypothesis of zero slope coefficients can be tested using a Chow F-statistic. This yields an observed value of 17.80 which is distributed as $F(10, 584)$ under the null hypothesis. Again, the null is soundly rejected. This F-test is in fact the BRMR test considered in section 13.6. As described in section 13.8, McFadden's R^2 is given by $R^2_5 = 1 - [\log\ell_u/\log\ell_r]$ which for the probit model yields

$$R^2_5 = 1 - (313.380/390.918) = 0.198.$$

For the logit model, McFadden's R^2_5 is 0.201.

Table 13.4 Probit Estimates: Union Participation

Dependent Variable:	UNION
Method:	ML – Binary Probit
Sample:	1 595
Included observations:	595
Convergence achieved after 5 iterations	
Covariance matrix computed using second derivatives	

Variable	Coefficient	Std. Error	z-Statistic	Prob.
EX	–0.006932	0.005745	–1.206491	0.2276
WKS	–0.060829	0.011785	–5.161666	0.0000
OCC	0.955490	0.152137	6.280476	0.0000
IND	0.092827	0.122774	0.756085	0.4496
SOUTH	–0.592739	0.139102	–4.261183	0.0000
SMSA	0.260700	0.128630	2.026741	0.0427
MS	0.350520	0.216284	1.620648	0.1051
FEM	–0.407026	0.277038	–1.469203	0.1418
ED	–0.057382	0.028842	–1.989515	0.0466
BLK	0.226482	0.228845	0.989675	0.3223
C	2.516784	0.762612	3.300217	0.0010

Mean dependent var	0.366387	S.D. dependent var	0.482222
S.E. of regression	0.420828	Akaike info criterion	1.090351
Sum squared resid	103.4242	Schwarz criterion	1.171484
Log likelihood	–313.3795	Hannan-Quinn criter.	1.121947
Restr. log likelihood	–390.9177	Avg. log likelihood	–0.526688
LR statistic (10 df)	155.0763	McFadden R-squared	0.198349
Probability(LR stat)	0.000000		
Obs with Dep=0	377	Total obs	595
Obs with Dep=1	218		

Table 13.5 Actual versus Predicted: Union Participation

	Predicted		Total
	Union = 0	Union = 1	
Actual			
Union = 0	OLS = 312 LOGIT = 316 Probit = 314	OLS = 65 LOGIT = 61 Probit = 63	377
Union = 1	OLS = 83 LOGIT = 82 Probit = 86	OLS = 135 LOGIT = 136 Probit = 132	218
Total	OLS = 395 LOGIT = 398 Probit = 400	OLS = 200 LOGIT = 197 Probit = 195	595

Table 13.6 Probit Estimates: Employment and Problem Drinking

Dependent Variable: EMPL
Method: ML – Binary Probit
Sample: 1 9822
Included observations: 9822
Convergence achieved after 7 iterations
QML (Huber/White) standard errors & covariance

Variable	Coefficient	Std. Error	z-Statistic	Prob.
90th Pctl.	−0.104947	0.058985	−1.779203	0.0752
UE88	−0.053277	0.014202	−3.751467	0.0002
AGE	0.099634	0.017118	5.820531	0.0000
AGE2	−0.001304	0.000205	−6.359957	0.0000
SCHOOLING	0.047183	0.006674	7.070238	0.0000
MARRIED	0.295292	0.054083	5.45998	0.0000
FAMILY SIZE	0.018891	0.014046	1.344943	0.1786
WHITE	0.394523	0.048336	8.162155	0.0000
EXCELLENT	1.816306	0.09834	18.46972	0.0000
VERY GOOD	1.778434	0.099148	17.93715	0.0000
GOOD	1.547836	0.098259	15.75266	0.0000
FAIR	1.043363	0.107722	9.685669	0.0000
NORTH EAST	0.034312	0.061999	0.553434	0.5800
MIDWEST	0.060491	0.053786	1.124662	0.2607
SOUTH	0.182121	0.054232	3.358186	0.0008
CENTER	−0.073053	0.051869	−1.408406	0.1590
OTHER MSA	0.075953	0.051307	1.480381	0.1388
QU1	−0.105484	0.05277	−1.998942	0.0456
QU2	−0.051323	0.052816	−0.971734	0.3312
QU3	−0.029342	0.054372	−0.539648	0.5894
CONSTANT	−3.017454	0.359214	−8.400162	0.0000

Mean dependent var	0.898188	S.D. dependent var	0.302417
S.E. of regression	0.277414	Akaike info criterion	0.553692
Sum squared resid	754.2706	Schwarz criterion	0.569069
Log likelihood	−2698.180	Hannan-Quinn criter.	0.558902
Avg. log likelihood	−0.274708		
Obs with Dep=0	1000	Total obs	9822
Obs with Dep=1	8822		

Example 2: Employment and Problem Drinking

Mullahy and Sindelar (1996) estimate a linear probability model relating employment and measures of problem drinking. The analysis is based on the 1988 Alcohol Supplement of the National Health Interview Survey. This regression was performed for Males and Females separately since the authors argue that women are less likely than men to be alcoholic, are more likely to abstain from consumption, and have lower mean alcohol consumption levels. They also report that women metabolize ethanol faster than do men and experience greater liver damage for the same level of consumption of ethanol. The dependent variable takes the value 1 if the individual was employed in the past two weeks and zero otherwise. The explanatory variables included

the 90th percentile of ethanol consumption in the sample (18 oz. for males and 10.8 oz. for females) and zero otherwise. The state unemployment rate in 1988 (UE88), Age, Age^2, schooling, married, family size, and white. Health status dummies indicating whether the individual's health was excellent, very good, fair. Region of residence, whether the individual resided in the northeast, midwest or south. Also, whether he or she resided in center city or other metropolitan statistical area (not center city). Three additional dummy variables were included for the quarters in which the survey was conducted. Details on the definitions of these variables are given in Table 1 of Mullahy and Sindelar (1996). Table 13.6 gives the probit results based on $n = 9822$ males. These results show a negative relationship between the 90th percentile alcohol variable and the probability of being employed, but this has a p-value of 0.075. Mullahy and Sindelar find that for both men and women, problem drinking results in reduced employment and increased unemployment. Problem 13 asks the reader to verify these results as well as those in the original article by Mullahy and Sindelar (1996).

13.10 Multinomial Choice Models

In many economic situations, the choice may be among m alternatives where $m > 2$. These may be unordered alternatives like the selection of a mode of transportation, bus, car or train, or an occupational choice like lawyer, carpenter, teacher, etc., or they may be ordered alternatives like bond ratings, or the response to an opinion survey, which could vary from strongly agree to strongly disagree. Ordered response multinomial models utilize the extra information implicit in the ordinal nature of the dependent variable. Therefore, these models have a different likelihood than unordered response multinomial models and have to be treated separately.

13.10.1 Ordered Response Models

Suppose there are three bond ratings, A, AA and AAA. We sample n bonds and the i-th bond is rated A (which we record as $y_i = 0$) if its performance index $I_i^* < 0$, where 0 is again not restrictive. $I_i^* = x_i'\beta + u_i$, so the probability of an A rating or the $\Pr[y_i = 0]$ is

$$\pi_{1i} = \Pr[y_i = 0] = P[I_i^* < 0] = P[u_i < -x_i'\beta] = F(-x_i'\beta) \tag{13.34}$$

The i-th bond is rated AA (which we record as $y_i = 1$) if its performance index I_i^* is between 0 and c where c is a positive number, with probability

$$\begin{aligned} \pi_{2i} &= \Pr[y_i = 1] = P[0 \le I_i^* < c] \\ &= P[0 \le x_i'\beta + u_i < c] = F(c - x_i'\beta) - F(-x_i'\beta) \end{aligned} \tag{13.35}$$

The i-th bond is rated AAA (which we record as $y_i = 2$) if $I_i^* \ge c$, with probability

$$\pi_{3i} = \Pr[y_i = 2] = P[I_i^* \ge c] = P[x_i'\beta + u_i \ge c] = 1 - F(c - x_i'\beta) \tag{13.36}$$

F can be the logit or probit function. The log-likelihood function for the ordered probit is given by

$$\begin{aligned} \log \ell(\beta, c) &= \textstyle\sum_{y_i=0} \log(\Phi(-x_i'\beta)) + \sum_{y_i=1} \log[\Phi(c - x_i'\beta) - \Phi(-x_i'\beta)] \\ &\quad + \textstyle\sum_{y_i=2} \log[1 - \Phi(c - x_i'\beta)]. \end{aligned} \tag{13.37}$$

For the probabilities given in (13.34), (13.35) and (13.36), the marginal effects of changes in the regressors are:

$$\partial\pi_{1i}/\partial x_i = -f(-x_i'\beta)\beta \tag{13.38}$$
$$\partial\pi_{2i}/\partial x_i = [f(-x_i'\beta) - f(c - x_i'\beta)]\beta \tag{13.39}$$
$$\partial\pi_{3i}/\partial x_i = f(c - x_i'\beta)\beta \tag{13.40}$$

Generalizing this model to m bond ratings is straight forward. The likelihood, the score and the Hessian for the m-ordered probit model are given in Maddala (1983, pp. 47-49).

13.10.2 Unordered Response Models

There are m choices each with probability $\pi_{i1}, \pi_{i2}, \ldots, \pi_{im}$ for individual i. $y_{ij} = 1$ if individual i chooses alternative j, otherwise it is 0. This means that $\sum_{j=1}^m y_{ij} = 1$ and $\sum_{j=1}^m \pi_{ij} = 1$. The likelihood function for n individuals is a multinomial given by:

$$\ell = \prod_{i=1}^n (\pi_{i1})^{y_{i1}} (\pi_{i2})^{y_{i2}} .. (\pi_{im})^{y_{im}} \tag{13.41}$$

This model can be motivated by a utility maximization story where the utility that individual i derives from say the occupational choice j is denoted by U_{ij} and is a function of the job attributes for the i-th individual, i.e., some x_{ij}'s like the present value of potential earnings, and training cost/net worth for that job choice for individual i, see Boskin (1974).

$$U_{ij} = x_{ij}'\beta + \epsilon_{ij} \tag{13.42}$$

where β is a vector of implicit prices for these occupational characteristics. Therefore, the probability of choosing the first occupation is given by:

$$\begin{aligned} \pi_{i1} &= \Pr[U_{i1} > U_{i2}, U_{i1} > U_{i3}, \ldots, U_{i1} > U_{im}] \\ &= \Pr[\epsilon_{i2} - \epsilon_{i1} < (x_{i1}' - x_{i2}')\beta, \epsilon_{i3} - \epsilon_{i1} \\ &< (x_{i1}' - x_{i3}')\beta, \ldots, \epsilon_{im} - \epsilon_{i1} < (x_{i1}' - x_{im}')\beta] \end{aligned} \tag{13.43}$$

The normality assumption involves a number of integrals but has the advantage of not necessarily assuming the ϵ's to be independent. The more popular assumption computationally is the multinomial logit model. This arises if and only if the ϵ's are independent and identically distributed as a Weibull density function, see McFadden (1974). The latter is given by $F(z) = \exp(-\exp(-z))$. The difference between any two random variables with a Weibull distribution has a logistic distribution $\Lambda(z) = e^z/1 + e^z$, giving the *conditional logit model*:

$$\begin{aligned} \pi_{ij} &= \Pr[y_i = j] = \exp[(x_{ij} - x_{im})'\beta]/\{1 + \textstyle\sum_{j=1}^{m-1} \exp[(x_{ij} - x_{im})'\beta]\} \\ &= \exp[x_{ij}'\beta]/\textstyle\sum_{j=1}^m \exp[x_{ij}'\beta] \qquad \text{for} \qquad j = 1, 2, \ldots, m-1 \end{aligned} \tag{13.44}$$

and $\pi_{im} = \Pr[y_i = m] = 1/\{1 + \sum_{j=1}^{m-1} \exp[(x_{ij} - x_{im})'\beta]\} = \exp[x_{im}'\beta]/\sum_{j=1}^m \exp[x_{ij}'\beta]$. There are two consequences of this conditional logit specification. The first is that the odds of any two alternative occupations, say 1 and 2 is given by

$$\pi_{i1}/\pi_{i2} = \exp[(x_{i1} - x_{i2})'\beta]$$

and this does not change when the number of alternatives change from m to m^*, since the denominators divide out. Therefore, the odds are unaffected by an additional alternative. This

is known as the *independence of irrelevant alternatives* property and can represent a serious weakness in the conditional logit model. For example, suppose the choices are between a pony and a bicycle, and children choose a pony two-thirds of the time. Suppose that an additional alternative is made available, an additional bicycle but of a different color, then one would still expect two-thirds of the children to choose the pony and the remaining one-third to split choices among the bicycles according to their color preference. In the conditional logit model, however, the proportion choosing the pony must fall to one half if the odds relative to either bicycle is to remain two to one in favor of the pony. This illustrates the point that when two or more of the m alternatives are close substitutes, the conditional logit model may not produce reasonable results. This feature is a consequence of assuming the errors ϵ_{ij}'s as independent. Hausman and McFadden (1984) proposed a Hausman type test to check for the independence of these errors. They suggest that if a subset of the choices is truly irrelevant then omitting it from the model altogether will not change the parameter estimates systematically. Including them if they are irrelevant preserves consistency but is inefficient. The test statistic is

$$q = (\widehat{\beta}_s - \widehat{\beta}_f)'[\widehat{V}_s - \widehat{V}_f]^{-1}(\widehat{\beta}_s - \widehat{\beta}_f) \tag{13.45}$$

where s indicates the estimators based on the restricted subset and f denotes the estimator based on the full set of choices. This is asymptotically distributed as χ^2_k, where k is the dimension of β.

Second, in this specification, none of the x_{ij}'s can be constant across different alternatives, because the corresponding β will not be identified. This means that we cannot include individual specific variables that do not vary across alternatives like race, sex, age, experience, income, etc. The latter type of data is more frequent in economics, see Schmidt and Strauss (1975). In this case the specification can be modified to allow for a differential impact of the explanatory variables upon the odds of choosing one alternative rather than the other:

$$\pi_{ij} = \Pr[y_i = j] = \exp(x'_{ij}\beta_j)/\textstyle\sum_{j=1}^{m}\exp(x'_{ij}\beta_j) \quad \text{for} \quad j = 1, \ldots, m \tag{13.46}$$

where now the parameter vector is indexed by j. If the x_{ij}'s are the same for every j, then

$$\pi_{ij} = Pr[y_i = j] = \exp(x'_i\beta_j)/\textstyle\sum_{j=1}^{m}\exp(x'_i\beta_j) \quad \text{for} \quad j = 1, \ldots, m \tag{13.47}$$

This is the model used by Schmidt and Strauss (1975). A normalization would be to take $\beta_m = 0$, in which case, we get the *multinomial logit* model

$$\pi_{im} = 1/\textstyle\sum_{j=1}^{m}\exp(x'_i\beta_j) \tag{13.48}$$

and

$$\pi_{ij} = \exp(x'_i\beta_j)/[1 + \textstyle\sum_{j=1}^{m-1}\exp(x'_i\beta_j)] \quad \text{for} \quad j = 1, 2, \ldots, m-1. \tag{13.49}$$

The likelihood function, score equations, Hessian and information matrices are given in Maddala (1983, pp. 36-37).

13.11 The Censored Regression Model

Suppose one is interested in the amount one is willing to spend on the purchase of a durable good. For example, a car. In this case, one would observe the expenditures only if the car is

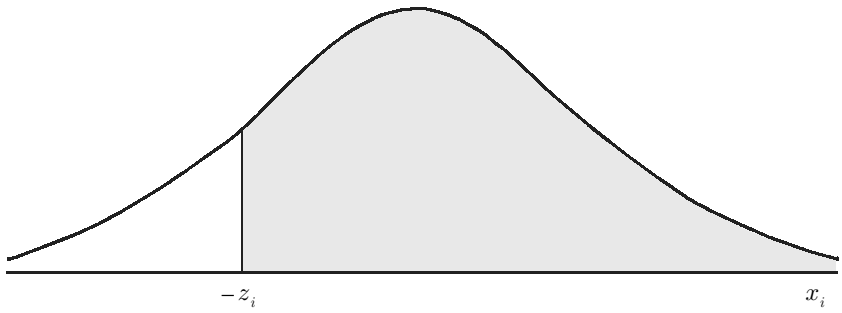


Figure 13.2 Truncated Normal Distribution

bought, so

$$y_i^* = x_i'\beta + u_i \quad \text{if} \quad y_i^* > 0 \tag{13.50}$$

where x_i denotes a vector of household characteristics, such as income, number of children or education. y_i^* is a latent variable, in this case the amount one is willing to spend on a car. We observe $y_i = y_i^*$ only if $y_i^* > 0$ and we set $y_i = 0$ if $y_i^* \leq 0$. The censoring at zero is of course arbitrary, and the u_i's are assumed to be $\text{IIN}(0, \sigma^2)$. This is known as the *Tobit* model after Tobin (1958). In this case, we have *censored* observations since we do not observe any y^* that is negative. All we observe is the fact that this household did not buy a car and a corresponding vector x_i of this household's characteristics. Without loss of generality, we assume that the first n_1 observations have positive y_i^*'s and the remaining $n_0 = n - n_1$ observations have non-positive y_i^*'s. In this case, OLS on the first n_1 observations, i.e., using only the positive observed y_i^*'s would be biased since u_i does not have zero mean. In fact, by omitting observations for which $y_i^* \leq 0$ from the sample, one is only considering disturbances from (13.50) such that $u_i > -x_i'\beta$. The distribution of these u_i's is a truncated normal density given in Figure 13.2. The mean of this density is not zero and is dependent on β, σ^2 and x_i. More formally, the regression function can be written as:

$$\begin{aligned} E(y_i^*/x_i, y_i^* > 0) &= x_i'\beta + E[u_i/y_i^* > 0] = x_i'\beta + E[u_i/u_i > -x_i'\beta] \\ &= x_i'\beta + \sigma\gamma_i \quad \text{for} \quad i = 1, 2, \ldots, n_1 \end{aligned} \tag{13.51}$$

where $\gamma_i = \phi(-z_i)/[1 - \Phi(-z_i)]$ and $z_i = x_i'\beta/\sigma$. See Greene (1993, p. 685) for the moments of a truncated normal density or the Appendix to this chapter. OLS on the positive y_i^*'s omits the second term in (13.51), and is therefore biased and inconsistent.

A simple two-step can be used to estimate (13.51). First, we define a dummy variable d_i which takes the value 1 if y_i^* is observed and 0 otherwise. This allows us to perform probit estimation on the whole sample, and provides us with a consistent estimator of (β/σ). Also, $P[d_i = 1] = P[y_i^* > 0] = P[u_i > -x_i'\beta]$ and $P[d_i = 0] = P[y_i^* \leq 0] = P[u_i \leq -x_i'\beta]$. Therefore, the likelihood function is given by

$$\begin{aligned} \ell &= \textstyle\prod_{i=1}^n [P(u_i \leq -x_i'\beta)]^{1-d_i} [P(u_i > -x_i'\beta)]^{d_i} \\ &= \textstyle\prod_{i=1}^n \Phi(z_i)^{d_i} [1 - \Phi(z_i)]^{1-d_i} \quad \text{where} \quad z_i = x_i'\beta/\sigma \end{aligned} \tag{13.52}$$

and once β/σ is estimated, we substitute these estimates in z_i and γ_i given below (13.51) to get $\widehat{\gamma}_i$. The second step is to estimate (13.51) using only the positive y_i^*'s with $\widehat{\gamma}_i$ substituted for γ_i. The resulting estimator of β is consistent and asymptotically normal, see Heckman (1976, 1979).

Alternatively, one can use maximum likelihood procedures to estimate the Tobit model. Note that we have two sets of observations: (i) the positive y_i^*'s with $y_i = y_i^*$, for which we can write the density function $N(x_i'\beta, \sigma^2)$, and (ii) the non-positive y_i^*'s for which we assign $y_i = 0$ with probability

$$\Pr[y_i = 0] = \Pr[y_i^* < 0] = \Pr[u_i < -x_i'\beta] = \Phi(-x_i'\beta/\sigma) = 1 - \Phi(x_i'\beta/\sigma) \tag{13.53}$$

The probability over the entire censored region gets assigned to the censoring point. This allows us to write the following log-likelihood:

$$\begin{aligned} \log\ell &= -(1/2)\textstyle\sum_{i=1}^{n_1} \log(2\pi\sigma^2) - (1/2\sigma^2)\sum_{i=1}^{n_1}(y_i - x_i'\beta)^2 \\ &\quad + \textstyle\sum_{i=n_1+1}^{n} \log[1 - \Phi(x_i'\beta/\sigma)] \end{aligned} \tag{13.54}$$

Differentiating with respect to β and σ^2, see Maddala (1983, p. 153), one gets

$$\partial\log\ell/\partial\beta = \textstyle\sum_{i=1}^{n_1}(y_i - x_i'\beta)x_i/\sigma^2 - \sum_{i=n_1+1}^{n} \phi_i x_i/\sigma[1 - \Phi_i] \tag{13.55}$$

$$\partial\log\ell/\partial\sigma^2 = \textstyle\sum_{i=1}^{n_1}(y_i - x_i'\beta)^2/2\sigma^4 - (n_1/2\sigma^2) + \sum_{i=n_1+1}^{n} \phi_i x_i'\beta/[2\sigma^3(1 - \Phi_i)] \tag{13.56}$$

where Φ_i and ϕ_i are evaluated at $z_i = x_i'\beta/\sigma$.

Premultiplying (13.55) by $\beta'/2\sigma^2$ and adding the result to (13.56), one gets

$$\widehat{\sigma}^2_{MLE} = \textstyle\sum_{i=1}^{n_1}(y_i - x_i'\beta)y_i/n_1 = Y_1'(Y_1 - X_1\beta)/n_1 \tag{13.57}$$

where Y_1 denotes the $n_1 \times 1$ vector of non-zero observations on y_i, X_1 is the $n_1 \times k$ matrix of values of x_i for the non-zero y_i's. Also, after multiplying throughout by σ, (13.55) can be written as:

$$-X_0'\gamma_0 + X_1'(Y_1 - X_1\beta)/\sigma = 0 \tag{13.58}$$

where X_0 denotes the $n_0 \times k$ matrix of x_i's for which y_i is zero, γ_0 is an $n_0 \times 1$ vector of γ_i's $= \phi_i/[1 - \Phi_i]$ evaluated at $z_i = x_i'\beta/\sigma$ for the observations for which $y_i = 0$. Solving (13.58) one gets

$$\widehat{\beta}_{MLE} = (X_1'X_1)^{-1}X_1'Y_1 - \sigma(X_1'X_1)^{-1}X_0'\gamma_0 \tag{13.59}$$

Note that the first term in (13.59) is the OLS estimator for the first n_1 observations for which y_i^* is positive.

One can use the Newton-Raphson procedure or the method of scoring, for the second derivatives of the log-likelihood, see Maddala (1983, pp. 154-156). These can be easily computed with the tobit command in Stata. Note that for the Tobit specification, both β and σ^2 are identified. This is contrasted to the logit and probit specifications where only the ratio (β/σ^2) is identified. Maddala warns that the Tobit specification is not necessarily the right specification every time we have zero observations. It is applicable only in those cases where the latent variable can, in principle, take negative values and the observed zero values are a consequence of censoring and non-observability. In fact, one cannot have negative expenditures on a car, negative hours of work or negative wages. However, one can enter employment and earn wages when one's observed wage is larger than the reservation wage. Let y^* be the difference between observed wage and reservation wage. Only if y^* is positive will wages be observed.

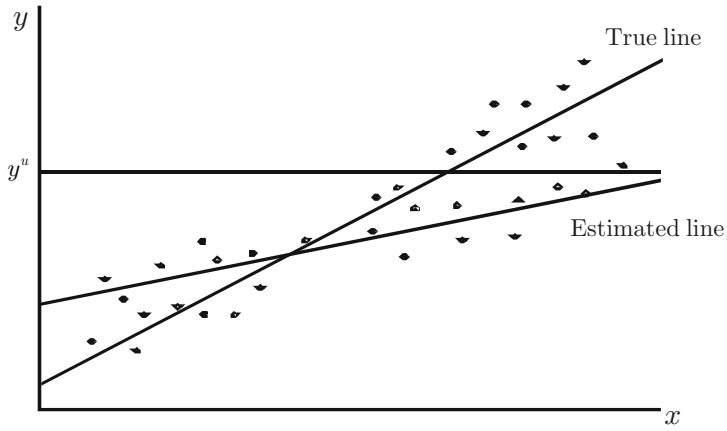


Figure 13.3 Truncated Regression Model

13.12 The Truncated Regression Model

The truncated regression model excludes or truncates some observations from the sample. For example, in studying poverty we exclude the rich, say with earnings larger than some upper limit y^u from our sample. The sample is therefore not random and applying least squares to the truncated sample lead to biased and inconsistent results, see Figure 13.3. This differs from censoring. In the latter case, no data is excluded. In fact, we observe the characteristics of all households even those that do not actually purchase a car. The truncated regression model is given by

$$y_i^* = x_i'\beta + u_i \quad i = 1, 2, \ldots, n \quad \text{with} \quad u_i \sim \text{IIN}(0, \sigma^2) \tag{13.60}$$

where y_i^* is for example earnings of the i-th household and x_i contains determinants of earnings like education, experience, etc. The sample contains observations on individuals with $y_i^* \leq y^u$. The probability that y_i^* will be observed is

$$\Pr[y_i^* \leq y^u] = \Pr[x_i'\beta + u_i \leq y^u] = \Pr[u_i < y^u - x_i'\beta] = \Phi(\frac{1}{\sigma}(y^u - x_i'\beta)) \tag{13.61}$$

In addition, using the results of a truncated normal density, see Greene (1993, p. 685)

$$E(u_i/y_i^* \leq y^u) = \frac{-\sigma\phi((y^u - x_i'\beta)/\sigma)}{\Phi((y^u - x_i'\beta)/\sigma)} \tag{13.62}$$

which is not necessarily zero. From (13.60) one can see that $E(y_i^*/y_i^* \leq y^u) = x_i'\beta + E(u_i/y_i^* < y^u)$. Therefore, OLS on (13.60) using the observed y_i^* is biased and inconsistent because it ignores the term in (13.62).

The density of y_i^* is normal but its total area is given by (13.61). A proper density function has to have an area of 1. Therefore, the density of y_i^* conditional on $y_i^* \leq y^u$ is simply the conditional density of y_i^* restricted to values of $y_i^* \leq y^u$ divided by the $\Pr[y_i^* \leq y^u]$, see the Appendix to this chapter:

$$\begin{aligned} f(y_i^*) &= \frac{\phi((y_i^* - x_i'\beta/\sigma)}{\sigma\Phi((y^u - x_i'\beta)/\sigma)} \quad \text{if} \quad y_i^* \leq y^u \\ &= 0 \qquad \text{otherwise} \end{aligned} \tag{13.63}$$

The log-likelihood function is therefore

$$\begin{aligned} \log\ell &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\textstyle\sum_{i=1}^{n}(y_i^* - x_i'\beta)^2 \\ &\quad - \textstyle\sum_{i=1}^{n}\log\Phi\left(\dfrac{y^u - x_i'\beta}{\sigma}\right) \end{aligned} \tag{13.64}$$

It is the last term which makes MLE differ from OLS on the observed sample. Hausman and Wise (1977) applied the truncated regression model to data from the New Jersey negative-income-tax experiment where families with incomes higher than 1.5 times the 1967 poverty line were eliminated from the sample.

13.13 Sample Selectivity

In labor economics, one observes the market wages of individuals only if the worker participates in the labor force. This happens when the worker's market wage exceeds his or her reservation wage. In a study of earnings, one does not observe the reservation wage and for non-participants in the labor force we record a zero market wage. This sample is censored because we observe the characteristics of these non-labor participants. If we restrict our attention to labor market participants only, then the sample is truncated. This example needs special attention because the censoring is not based directly on the dependent variable, as in section 13.11. Rather, it is based on the difference between market wage and reservation wage. This latent variable which determines the sample selection is correlated with the dependent variable. Hence, least squares on this model results in *selectivity bias*, see Heckman (1976, 1979). A sample generated by this type of self-selection may not represent the true population distribution no matter how big the sample size. However, one can correct for self-selection bias if the underlying sampling generating process can be understood and relevant identification conditions are available, see Lee (2001) for a excellent summary and the references cited there. In order to demonstrate this, let the earnings equation be given by

$$w_i^* = x_{1i}'\beta + u_i \qquad i = 1, 2, \ldots, n \tag{13.65}$$

and the labor participation (or selection) equation be given by

$$y_i^* = x_{2i}'\gamma + v_i \qquad i = 1, 2, \ldots, n \tag{13.66}$$

where u_i and v_i are bivariate normally distributed with mean zero and variance-covariance

$$\text{var}\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{bmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{bmatrix} \tag{13.67}$$

Normalizing the variance of v_i to be 1 is not restrictive, since only the sign of y_i^* is observed. In fact, we only observe w_i and y_i where

$$\begin{aligned} w_i &= w_i^* \quad \text{if} \quad y_i^* > 0 \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{13.68}$$

and

$$\begin{aligned} y_i &= 1 \quad \text{if} \quad y_i^* > 0 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

We observe $(y_i = 0, w_i = 0)$ and $(y_i = 1, w_i = w_i^*)$ only. The log-likelihood for this model is

$$\sum_{y_i=0} \log \Pr[y_i = 0] + \sum_{y_i=1} \log \Pr[y_i = 1] f(w_i^*/y_i = 1) \tag{13.69}$$

where $f(w_i^*/y_i = 1)$ is the conditional density of w_i^* given that $y_i = 1$. The second term can also be written as $\Sigma_{y_i=1} \log \Pr[y_i = 1/w_i^*] f(w_i^*)$ which is another way of factoring the joint density function. $f(w_i^*)$ is in fact a normal density with conditional mean $x_{1i}'\beta$ and variance σ^2. Using properties of the bivariate normal density, one can write

$$y_i^* = x_{2i}'\gamma + \rho \left(\frac{1}{\sigma}(w_i^* - x_{1i}'\beta) \right) + \epsilon_i \tag{13.70}$$

where $\epsilon_i \sim \text{IIN}(0, \sigma^2(1 - \rho^2))$. Therefore,

$$\Pr[y_i = 1] = \Pr[y_i^* > 0] = \Phi \left(\frac{x_{2i}'\gamma + \rho((w_i - x_{1i}'\beta)/\sigma)}{\sqrt{1 - \rho^2}} \right) \tag{13.71}$$

where w_i has been substituted for w_i^* since $y_i = 1$. The likelihood function in (13.69) becomes

$$\sum_{y_i=0} \log(\Phi(-x_{2i}'\gamma)) + \sum_{y_i=1} \log \left(\frac{1}{\sigma} \phi(w_i - x_{1i}'\beta) \right) \tag{13.72}$$

$$+ \sum_{y_i=1} \log \Phi \left(\frac{x_{2i}'\gamma + \rho((w_i - x_{1i}'\beta)/\sigma}{\sqrt{1 - \rho^2}} \right)$$

MLE may be computationally burdensome. Heckman (1976) suggested a two-step procedure which is based on rewriting (13.65) as

$$w_i^* = \beta x_{1i}' + \rho\sigma v_i + \eta_i \tag{13.73}$$

and replacing w_i^* by w_i and v_i by its conditional mean $E[v_i/y_i = 1]$. Using the results on truncated density, this conditional mean is given by $\phi(x_{2i}'\gamma)/\Phi(-x_{2i}'\gamma)$ known also as the *inverse Mills ratio*. Hence, (13.73) becomes

$$w_i = x_{1i}'\beta + \rho\sigma \frac{\phi(x_{2i}'\gamma)}{\Phi(-x_{2i}'\gamma)} + \text{residual} \tag{13.74}$$

Heckman's (1976) two-step estimator consists of (i) running a probit on (13.66) in the first step to get a consistent estimate of γ, (ii) substituting the estimated inverse Mills ratio in (13.74) and running OLS. Since σ is positive, this second stage regression provides a test for sample selectivity, i.e., for $\rho = 0$, by checking whether the t-statistic on the estimated inverse Mills ratio is significant. This statistic is asymptotically distributed as $N(0, 1)$. Rejecting H_o implies there is a selectivity problem and one should not rely on OLS on (13.65) which ignores the selectivity bias term in (13.74). Davidson and MacKinnon (1993) suggest performing MLE using (13.72) rather than relying on the two-step results in (13.74) if the former is not computationally burdensome. Note that the Tobit model for car purchases given in (13.50) can be thought of as a special case of the sample selectivity model given by (13.65) and (13.66). In fact, the Tobit model assumes that the selection equation (the decision to buy a car) and the car expenditure equation (conditional on the decision to buy) are identical. Therefore, if one thinks that the specification of the selection equation is different from that of the expenditure equation, then

one should *not* use the Tobit model. Instead, one should proceed with the two equation sample selectivity model discussed in this section. It is also important to emphasize that for the censored, truncated and sample selectivity models, normality and homoskedasticity are crucial assumptions. Suggested tests for these assumptions are given in Bera, Jarque and Lee (1984), Lee and Maddala (1985) and Pagan and Vella (1989). Alternative estimation methods that are more robust to violations of normality and heteroskedasticity include symmetrically trimmed least squares for Tobit models and least absolute deviations estimation for censored regression models. These were suggested by Powell (1984, 1986).

Notes

1. This is based on Davidson and MacKinnon (1993, pp. 523-526).
2. A binary response model attempts to explain a zero-one (or binary) dependent variable.
3. One should not use nR^2 as the test statistic because the total sum of squares in this case is not n.

Problems

1. *The Linear Probability Model.*
   (a) For the linear probability model described in (13.1), show that for $E(u_i)$ to equal zero, we must have $\Pr[y_i = 1] = x_i'\beta$.
   (b) Show that u_i is heteroskedastic with $\text{var}(u_i) = x_i'\beta(1 - x_i'\beta)$.
2. Consider the general log-likelihood function given in (13.16). Assume that all the regression slopes are zero except for the constant α.
   (a) Show that maximizing $\log\ell$ with respect to the constant α yields $\widehat{F}(\alpha) = \bar{y}$, where $\bar{y}$ is the proportion of the sample with $y_i = 1$.
   (b) Conclude that the value of the maximized likelihood is $\log\ell_r = n[\bar{y}\log\bar{y} + (1 - \bar{y})\log(1 - \bar{y})]$.
   (c) Verify that for the union participation example in section 13.9 that $\log\ell_r = -390.918$.
3. For the union participation example considered in section 13.9:
   (a) Replicate Tables 13.3 and 13.5.
   (b) Using the measures of fit considered in section 13.8, compute $R_1^2, R_2^2, \ldots, R_5^2$ for the logit and probit models.
   (c) Compute the predicted value for the 10th observation using OLS, logit and probit models. Also, the corresponding standard errors.
   (d) The industry variable (IND) was not significant in all models. Drop this variable and run OLS, logit and probit. How do the results change? Compare with Tables 13.3 and 13.5.
   (e) Using the model results in part (d), test that the slope coefficients are all zero for the logit, probit, and linear probability models.
   (f) Test that the coefficients of IND, FEM and BLK in Table 13.3 are jointly insignificant using a LR test, a Wald test and a BRMR using OLS, logit and probit.

4. For the data used in the union participation example in section 13.9:

   (a) Run OLS, logit and probit using as the dependent variable OCC which is one if the individual is in a blue-collar occupation, and zero otherwise. For the independent variables use ED, WKS, EXP, SOUTH, SMSA, IND, MS, FEM and BLK. Compare the coefficient estimates across the three models. What variables are significant?

   (b) Using the measures of fit considered in section 13.8, compute $R_1^2, R_2^2, \ldots, R_5^2$ for the logit and probit models.

   (c) Tabulate the actual versus predicted values for OCC from all three model results, like Table 13.5 for Union. What is the proportion of correct decisions for OLS, logit and probit?

   (d) Test that the slope coefficients are all zero for the logit, probit, and linear probability models.

5. *Truncated Uniform Density.* Let x be a uniformly distributed random variable with density

$$f(x) = \frac{1}{2} \quad \text{for} \quad -1 < x < 1$$

   (a) What is the density function of $f\,(x/x > -1/2)$? **Hint:** Use the definition of a conditional density

$$f\,(x/x > -1/2) = f(x)/\Pr[x > -1/2] \quad \text{for} \quad -\frac{1}{2} < x < 1.$$

   (b) What is the conditional mean $\mathrm{E}(x/x > -1/2)$? How does it compare with the unconditional mean of x? Note that because we truncated the density from below, the new mean should shift to the right.

   (c) What is the conditional variance $\mathrm{var}(x/x > -1/2)$? How does it compare to the unconditional $\mathrm{var}(x)$? (Truncation reduces the variance).

6. *Truncated Normal Density.* Let x be $N(1,1)$. Using the results in the Appendix, show that:

   (a) The conditional density $f(x/x > 1) = 2\phi(x-1)$ for $x > 1$ and $f(x/x < 1) = 2\phi(x-1)$ for $x < 1$.

   (b) The conditional mean $E(x/x > 1) = 1 + 2\phi(0)$ and $E(x/x < 1) = 1 - 2\phi(0)$. Compare with the unconditional mean of x.

   (c) The conditional variance $\mathrm{var}(x/x > 1) = \mathrm{var}(x/x < 1) = 1 - 4\phi^2(0)$. Compare with the unconditional variance of x.

7. *Censored Normal Distribution.* This is based on Greene (1993, pp. 692-693). Let y^* be $N(\mu, \sigma^2)$ and define $y = y^*$ if $y^* > c$ and $y = c$ if $y^* < c$ for some constant c.

   (a) Verify the $E(y)$ expression given in (A.7).

   (b) Derive the $\mathrm{var}(y)$ expression given in (A.8). **Hint:** Use the fact that

$$\mathrm{var}(y) = E(\text{conditional variance}) + \mathrm{var}(\text{conditional mean})$$

   and the formulas given in the Appendix for conditional and unconditional means of a truncated normal random variable.

   (c) For the special case of $c = 0$, show that (A.7) simplifies to $E(y) = \Phi(\mu/\sigma)\left[\mu + \dfrac{\sigma\phi(\mu/\sigma)}{\Phi(\mu/\sigma)}\right]$ and (A.8) simplifies to

$$\mathrm{var}(y) = \sigma^2\Phi\left(\frac{\mu}{\sigma}\right)\left[1 - \delta\left(\frac{-\mu}{\sigma}\right) + \left(-\frac{\mu}{\sigma} - \frac{\phi(/\sigma)}{\Phi(\mu/\sigma)}\right)^2 \Phi\left(-\frac{\mu}{\sigma}\right)\right]$$

where $\delta\left(\frac{-\mu}{\sigma}\right) = \frac{\phi(\mu/\sigma)}{\Phi(\mu/\sigma)}\left[\frac{\phi(\mu/\sigma)}{\Phi(\mu/\sigma)} + \frac{\mu}{\sigma}\right]$. Similar expressions can be derived for censoring of the upper part rather than the lower part of the distribution.

8. Dhillon, Shilling and Sirmans (1987) considered the economic decision of choosing between fixed and adjustable rate mortgages. The data consisted of 78 households borrowing money from a Louisiana mortgage banker. Of these, 46 selected fixed rate mortgages and 32 selected uncapped adjustable rate mortgages. This data set can be downloaded from the Springer web site and is labelled DHILLON.ASC. It was obtained from Lott and Ray (1992). These variables include:

Y	= 0 if adjustable rate and 1 if fixed rate.
BA	= Age of the borrower.
BS	= Years of schooling for the borrower.
NW	= Net worth of the borrower.
FI	= Fixed interest rate.
PTS	= Ratio of points paid on adjustable to fixed rate mortgages.
MAT	= Ratio of maturities on adjustable to fixed rate mortgages.
MOB	= Years at the present address.
MC	= 1 if the borrower is married and 0 otherwise.
FTB	= 1 if the borrower is a first-time home buyer and 0 otherwise.
SE	= 1 if the borrower is self-employed and 0 otherwise.
YLD	= The difference between the 10-year treasury rate less the 1-year treasury rate.
MARG	= The margin on the adjustable rate mortgage.
CB	= 1 if there is a co-borrower and 0 otherwise.
STL	= Short-term liabilities.
LA	= Liquid assets.

The probability of choosing a variable rate mortgage is a function of personal borrower characteristics as well as mortgage characteristics. The efficient market hypothesis state that only cost variables and not personal borrower characteristics influence the borrower's decision between fixed and adjustable rate mortgages. Cost variables include FI, MARG, YLD, PTS and MAT. The rest of the variables are personal characteristics variables. The principal agent theory suggests that information is asymmetric between lender and borrower. Therefore, one implication of this theory is that the personal characteristics of the borrower will be significant in the choice of mortgage loan.

(a) Run OLS of Y on all variables in the data set. For this linear probability model what does the F-statistic for the significance of all slope coefficients yield? What is the R^2? How many predictions are less than zero or larger than one?

(b) Using only the cost variables in the restricted regression, test that personal characteristics are jointly insignificant. **Hint:** Use the Chow-F statistic. Do you find support for the efficient market hypothesis?

(c) Run the above model using the logit specification. Test the efficient market hypothesis. Does your conclusion change from that in part (b)? **Hint:** Use the likelihood ratio test or the BRMR.

(d) Do part (c) using the probit specification.

9. *Sampling Distribution of OLS Under a Logit Model.* This is based on Baltagi (2000).

Consider a simple logit regression model

$$y_t = \Lambda(\beta x_t) + u_t$$

for $t = 1, 2$, where $\Lambda(z) = e^z/(1 + e^z)$ for $-\infty < z < \infty$. Let $\beta = 1$, $x_1 = 1$, $x_2 = 2$ and assume that the u_t's are independent with mean zero.

(a) Derive the sampling distribution of the least squares estimator of β, i.e., assuming a linear probability model when the true model is a logit model.

(b) Derive the sampling distribution of the least squares residuals and verify the estimated variance of $\widehat{\beta}_{OLS}$ is biased.

10. *Sample Selection and Non-response.* This is based on Manski (1995), see the Appendix to this chapter. Suppose we are interested in estimating the probability than an individual who is homeless at a given date has a home six months later. Let $y = 1$ if the individual has a home six months later and $y = 0$ if the individual remains homeless. Let x be the sex of the individual and let $z = 1$ if the individual was located and interviewed and zero otherwise. 100 men and 31 women were initially interviewed. Six months later, only 64 men and 14 women were located and interviewed. Of the 64 men, 21 exited homelessness. Of the 14 women only 3 exited homelessness.

(a) Compute $\Pr[y = 1/\text{Male}, z = 1]$, $\Pr[z = 1/\text{Male}]$ and the bound on $\Pr[y = 1/\text{Male}]$.

(b) Compute $\Pr[y = 1/\text{Female}, z = 1]$, $\Pr[z = 1/\text{Female}]$ and the bound on $\Pr[y = 1/\text{Female}]$.

(c) Show that the width of the bound is equal to the probability of attrition. Which bound is tighter? Why?

11. *Does the Link Matter?* This is based on Santos Silva (1999). Consider a binary random variable Y_i such that

$$P(Y_i = 1|x) = F(\beta_0 + \beta_1 x_i), \qquad i = 1, \ldots, n,$$

where the link $F(\cdot)$ is a continuous distribution function.

(a) Write down the log-likelihood function and the first-order conditions of maximization with respect to β_0 and β_1.

(b) Consider the case where x_i only assumes two different values, without loss of generality, let it be 0 and 1. Show that $\widehat{F}(1) = \sum_{x_i=1} y_i/n_1$, where n_1 is the number of observations for which $x_i = 1$. Also, show that $\widehat{F}(0) = \sum_{x_i=0} y_i/(n - n_1)$.

(c) What are the maximum likelihood estimates of β_0 and β_1?

(d) Show that the value of the log-likelihood function evaluated at the maximum likelihood estimates of β_0 and β_1 is the same, independent of the form of the link function.

12. Ruhm (1996) considered the effect of beer taxes and a variety of alcohol-control policies on motor vehicle fatality rates, see section 13.4. The data is for 48 states (excluding Alaska, Hawaii and the District of Columbia) over the period 1982-1988. This data set can be downloaded from the Stock and Watson (2003) web site at www.aw.com/stock_watson. Using this data set replicate the results in Table 13.1.

13. Mullahy and Sindelar (1996) considered the effect of problem drinking on employment and unemployment. The data set is based on the 1988 Alcohol Supplement of the National Health Interview Survey. This can be downloaded from the *Journal of Applied Econometrics* web site at http://qed.econ.queensu.ca/jae/2002-v17.4/terza/.

(a) Replicate Table 13.6 and run also the logit and OLS regressions with robust White standard errors. The OLS results should match those given in Table 5 of Mullahy and Sindelar (1996).

(b) Mullahy and Sindelar (1996) performed similar regressions for females and for the dependent variable taking the value of 1 if the individual is unemployed and zero otherwise. Replicate the OLS results in Tables 5 and 6 of Mullahy and Sindelar (1996) and perform the corresponding logit and probit regressions. What is your conclusion on the relationship between problem drinking and unemployment?

14. Papke and Wooldridge (1996) studied the effect of match rates on participation rates in 401(K) pension plans. The data are from the 1987 IRS Form 5500 reports of pension plans with more than 100 participants. This data set can be downloaded from the *Journal of Applied Econometrics* web site at http://qed.econ.queensu.ca/jae/1996-V11.6/papke.

    (a) Replicate Tables I and II of Papke and Wooldridge (1996).

    (b) Run the specification tests (RESET) described in Papke and Wooldridge (1996).

    (c) Compare OLS and logit QMLE using R^2, specification tests, and predictions for various values of MRATE as done in Figure 1 of Papke and Wooldridge (1996, p. 630).

References

This chapter is based on the material in Hanushek and Jackson (1977), Maddala (1983), Davidson and MacKinnon (1993) and Greene (1993). Additional references include:

Amemiya, T. (1981), "Qualitative Response Models: A Survey," *Journal of Economic Literature*, 19: 1481-1536.

Amemiya, T. (1984), "Tobit Models: A Survey," *Journal of Econometrics*, 24: 3-61.

Baltagi, B.H. (2000), "Sampling Distribution of OLS Under a Logit Model," Problem 00.3.1, *Econometric Theory*, 16: 451.

Bera, A.K., C. Jarque and L.F. Lee (1984), "Testing the Normality Assumption in Limited Dependent Variable Models," *International Economic Review*, 25: 563-578.

Berkson, J. (1953), "A Statistically Precise and Relatively Simple Method of Estimating the Bio-Assay with Quantal Response, Based on the Logistic Function," *Journal of the American Statistical Association*, 48: 565-599.

Berndt, E., B. Hall, R. Hall and J. Hausman (1974), "Estimation and Inference in Nonlinear Structural Models," *Annals of Economic and Social Measurement*, 3/4: 653-665.

Boskin, M. (1974), "A Conditional Logit Model of Occupational Choice," *Journal of Political Economy*, 82: 389-398.

Cornwell, C. and P. Rupert (1988), "Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variables Estimators," *Journal of Applied Econometrics*, 3: 149-155.

Cox, D.R. (1970), *The Analysis of Binary Data* (Chapman and Hall: London).

Cragg, J. and R. Uhler (1970), "The Demand for Automobiles," *Canadian Journal of Economics,* 3: 386-406.

Davidson, R. and J. MacKinnon (1984), "Convenient Specification Tests for Logit and Probit Models," *Journal of Econometrics*, 25: 241-262.

Dhillon, U.S., J.D. Shilling and C.F. Sirmans (1987), "Choosing Between Fixed and Adjustable Rate Mortgages," *Journal of Money, Credit and Banking*, 19: 260-267.

Effron, B. (1978), "Regression and ANOVA with Zero-One Data: Measures of Residual Variation," *Journal of the American Statistical Association*, 73: 113-121.

Goldberger, A. (1964), *Econometric Theory* (Wiley: New York).

Gourieroux, C., A. Monfort and A. Trognon (1984), "Pseudo-Maximum Likelihood Methods: Theory," *Econometrica*, 52: 681-700.

Hanushek, E.A. and J.E. Jackson (1977), *Statistical Methods for Social Scientists* (Academic Press: New York).

Hausman, J. and D. McFadden (1984), "A Specification Test for Multinomial Logit Model," *Econometrica*, 52: 1219-1240.

Hausman, J.A. and D.A. Wise (1977), "Social Experimentation, Truncated Distributions, and Efficient Estimation," *Econometrica*, 45: 919-938.

Heckman, J. (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5: 475-492.

Heckman, J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47: 153-161.

Lee, L.F. (2001), "Self-Selection," Chapter 18 in B.H. Baltagi (ed.) *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Lee, L.F. and G.S. Maddala (1985), "The Common Structure of Tests for Selectivity Bias, Serial Correlation, Heteroskedasticity and Non-Normality in the Tobit Model," *International Economic Review*, 26: 1-20.

Lott, W.F. and S.C. Ray (1992), *Applied Econometrics: Problems with Data Sets* (The Dryden Press: New York).

Maddala, G. (1983), *Limited Dependent and Qualitative Variables in Econometrics* (Cambridge University Press: Cambridge).

Manski, C.F. (1995), *Identification Problems in the Social Sciences* (Harvard University Press: Cambridge).

McFadden, D. (1974), "The Measurement of Urban Travel Demand," *Journal of Public Economics*, 3: 303-328.

McCullagh, P. and J.A. Nelder (1989), *Generalized Linear Models* (Chapman and Hall: New York).

Mullahy, J. and J. Sindelar (1996), "Employment, Unemployment, and Problem Drinking," *Journal of Health Economics*, 15: 409-434.

Pagan, A.R. and F. Vella (1980), "Diagnostic Tests for Models Based on Individual Data: A Survey," *Journal of Applied Econometrics*, 4: S29-S59.

Papke, L.E. and J.M. Wooldridge (1996), "Econometric Methods for Fractional Response Variables with An Application to 401(K) Plan Participation Rates," *Journal of Applied Econometrics*, 11: 619-632.

Powell, J. (1984), "Least Absolute Deviations Estimation of the Censored Regression Model," *Journal of Econometrics*, 25: 303-325.

Powell, J. (1986), "Symmetrically Trimmed Least Squares Estimation for Tobit Models," *Econometrica*, 54: 1435-1460.

Pratt, J.W. (1981), "Concavity of the Log-Likelihood," *Journal of the American Statistical Association*, 76: 103-109.

Ruhm, C.J. (1996), "Alcohol Policies and Highway Vehicle Fatalities," *Journal of Health Economics*, 15: 435-454.

Schmidt, P. and R. Strauss (1975), "Estimation of Models With Jointly Dependent Qualitative Variables: A Simultaneous Logit Approach," *Econometrica*, 43: 745-755.

Wooldridge, J.M. (1991), "Specification Testing and Quasi-Maximum Likelihood Estimation," *Journal of Econometrics*, 48: 29-55.

Appendix

1. Truncated Normal Distribution

Let x be $N(\mu, \sigma^2)$, then for a constant c, the truncated density is given by

$$f(x/x > c) = \frac{f(x)}{\Pr[x > c]} = \frac{\frac{1}{\sigma}\phi([x-\mu]/\sigma)}{1 - \Phi\left(\frac{c-\mu}{\sigma}\right)} \quad c < x < \infty \tag{A.1}$$

where $\phi(z)$ denotes the p.d.f. and Φ denotes the c.d.f. of a $N(0,1)$ random variable. If the truncation is from above

$$f(x/x < c) = \frac{f(x)}{\Pr[x < c]} = \frac{\frac{1}{\sigma}\phi([x-\mu]/\sigma)}{\Phi\left(\frac{c-\mu}{\sigma}\right)} \quad -\infty < x < c \tag{A.2}$$

The conditional means are given by

$$E(x/x > c) = \mu + \sigma\frac{\phi(c^*)}{1 - \Phi(c^*)} \tag{A.3}$$

where $c^* = \dfrac{c-\mu}{\sigma}$, and

$$E(x/x < c) = \mu - \sigma\frac{\phi(c^*)}{\Phi(c^*)} \tag{A.4}$$

In other words, the truncated mean shifts to the right (left) if truncation is from below (above).

The conditional variances are given by $\sigma^2(1 - \delta(c^*))$ with $0 < \delta(c^*) < 1$ for all values of c^*.

$$\delta(c^*) = \frac{\phi(c^*)}{1 - \Phi(c^*)}\left[\frac{\phi(c^*)}{1 - \Phi(c^*)} - c^*\right] \quad \text{for} \quad x > c \tag{A.5}$$

$$= \frac{-\phi(c^*)}{\Phi(c^*)}\left[\frac{-\phi(c^*)}{\Phi(c^*)} - c^*\right] \quad \text{for} \quad x < c \tag{A.6}$$

In other words, the truncated variance is always less than the unconditional or untruncated variance. For more details, see Maddala (1983, p. 365) or Greene (1993, p. 685).

2. The Censored Normal Distribution

Let y^* be $N(\mu, \sigma^2)$, then for a constant c, define $y = y^*$ if $y^* > c$ and $y = c$ if $y^* < c$. Unlike the truncated normal density, the censored density assigns the entire probability of the censored region to the censoring point, i.e., $y = c$. So that $\Pr[y = c] = \Pr[y^* < c] = \Phi((c-\mu)/\sigma) = \Phi(c^*)$ where $c^* = (c-\mu)/\sigma$. For the uncensored region the probability of y^* remains the same and can be obtained from the normal density.

It is easy to show, see Greene (1993, p. 692) that

$$\begin{aligned} E(y) &= \Pr[y = c]E(y/y = c) + \Pr[y > c]E(y/y > c) \\ &= c\Phi(c^*) + (1 - \Phi(c^*))E(y^*/y^* > c) \\ &= c\Phi(c^*) + (1 - \Phi(c^*))\left[\mu + \sigma\frac{\phi(c^*)}{1 - \Phi(c^*)}\right] \end{aligned} \tag{A.7}$$

where $E(y^*/y^* > c)$ is obtained from the mean of a truncated normal density, see (A.3).

Similarly, one can show, see problem 7 or Greene (1993, p. 693) that

$$\text{var}(y) = \sigma^2 [1 - \Phi(c^*)] \left[1 - \delta(c^*) + \left(c^* - \frac{\phi(c^*)}{1 - \Phi(c^*)} \right)^2 \Phi(c^*) \right] \tag{A.8}$$

where $\delta(c^*)$ was defined in (A.5).

3. Sample Selection and Non-response

Non-response is a big problem plaguing survey data. Some individuals refuse to respond and some do not answer all the questions, especially on relevant economic variables like income. Suppose we interviewed randomly 150 individuals upon their graduation from high school. Among these, 50 were female and 100 were male. A year later, we try to re-interview these individuals to find out whether they are employed or not. Only 70 out of 100 males and 40 out of 50 females were located and interviewed a year later. Out of those re-interviewed, 60 males and 20 females were found to be employed. Let $y = 1$ if the individual is employed and zero if not. Let x be the sex of this individual and let $z = 1$ if this individual is located and interviewed a year later and zero otherwise.

Conditioning on sex of the respondent one can compute the probability of being employed a year after high school graduation as follows:

$$\Pr[y = 1/x] = \Pr[y = 1/x, z = 1] \Pr[z = 1/x] + \Pr[y = 1/x, z = 0] \Pr[z = 0/x]$$

In this case, $\Pr[y = 1/\text{Male}, z = 1] = 60/70$, $\Pr[z = 1/\text{Male}] = 70/100$ and $\Pr[z = 0/\text{Male}] = 30/100$. But the sampling process is uninormative about the non-respondents or the censored observations, i.e., $\Pr[y = 1/\text{Male}, z = 0]$. Therefore, in the absence of other information

$$\Pr[y = 1/\text{Male}] = (0.6) + (0.3) \Pr[y = 1/\text{Male}, z = 0]$$

Manski (1995) argues that one can estimate bounds on this probability. In fact, replacing $0 \leq \Pr[y = 1/\text{Male}, z = 0] \leq 1$ by its bounds, yields

$$0.6 \leq \Pr[y = 1/\text{Male}] \leq 0.9$$

with the width of the bound equal to the probability of non-response conditioning on Males, i.e., $\Pr[z = 0/\text{Male}] = 0.3$. Similarly, $0.4 \leq \Pr[y = 1/\text{Female}] \leq 0.6$ with the width of the bound equal to the probability of non-response conditioning on Females, i.e., $\Pr[z = 0/\text{Female}] = 10/50 = 0.2$. Manski (1995) argues that these bounds are informative and should be the starting point of empirical analysis. Researchers assuming that non-response is *ignorable* or *exogenous* are imposing the following restriction

$$\Pr[y = 1/\text{Male}, z = 1] = \Pr[y = 1/\text{Male}, z = 0] = \Pr[y = 1/\text{Male}] = 60/70$$
$$\Pr[y = 1/\text{Female}, z = 1] = \Pr[y = 1/\text{Female}, z = 0] = \Pr[y = 1/\text{Female}] = 20/40$$

To the extent that these probabilities are different casts doubt on the *ignorable* non-response assumption.

CHAPTER 14
Time-Series Analysis

14.1 Introduction

There has been an enormous amount of research in time-series econometrics, and many economics departments have required a time-series econometrics course in their graduate sequence. Obviously, one chapter on this topic will not do it justice. Therefore, this chapter will focus on some of the basic concepts needed for such a course. Section 14.2 defines what is meant by a *stationary* time-series, while sections 14.3 and 14.4 briefly review the Box-Jenkins and *Vector Autoregression* (VAR) methods for time-series analysis. Section 14.5 considers a random walk model and various tests for the existence of a *unit root*. Section 14.6 studies *spurious regressions* and *trend stationary* versus *difference stationary* models. Section 14.7 gives a simple explanation of the concept of *cointegration* and illustrates it with an economic example. Finally, section 14.8 looks at *Autoregressive Conditionally Heteroskedastic* (ARCH) time-series.

14.2 Stationarity

Figure 14.1 plots the consumption and personal disposable income data considered in Chapter 5. This was done using EViews. This is annual data from 1950 to 1993 expressed in real terms. Both series seem to be trending upwards over time. This may be an indication that these time-series are non-stationary. Having a time-series x_t that is trending upwards over time may invalidate all the standard asymptotic theory we have been relying upon in this book. In fact, $\sum_{t=1}^{T} x_t^2/T$ may not tend to a finite limit as $T \to \infty$ and using regressors such as x_t means that $X'X/T$ does not tend in probability limits to a finite positive definite matrix, see problem 6.

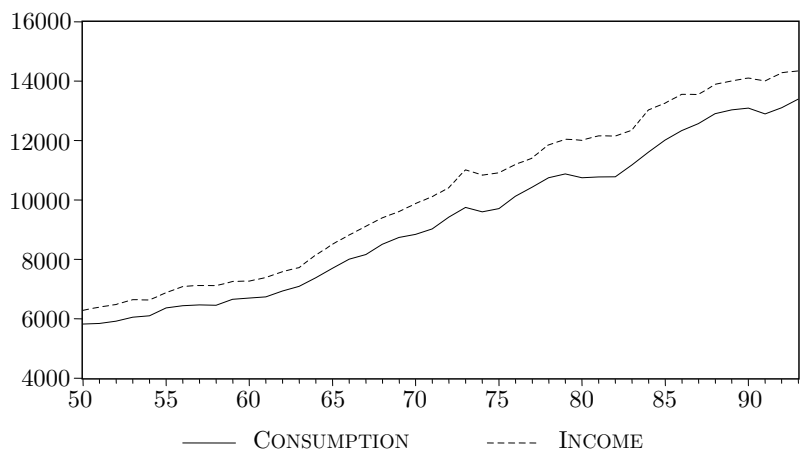


Figure 14.1 U.S. Consumption and Income, 1950–1993

Non-standard asymptotic theory will have to be applied which is beyond the scope of this book, see problem 8.

Definition: A time-series process x_t is said to be covariance stationary (or weakly stationary) if its mean and variance are constant and independent of time and the covariances given by $\text{cov}(x_t, x_{t-s}) = \gamma_s$ depend only upon the distance between the two time periods, but not the time periods per se.

In order to check the time-series for weak stationarity one can compute its *autocorrelation function*. This is given by ρ_s= correlation $(x_t, x_{t-s}) = \gamma_s/\gamma_o$. These are correlation coefficients taking values between -1 and $+1$.

The sample counterparts of the variance and covariances are given by

$$\widehat{\gamma}_o = \sum_{t=1}^{T}(x_t - \bar{x})^2/T$$

$$\widehat{\gamma}_s = \sum_{t=1}^{T-s}(x_t - \bar{x})(x_{t+s} - \bar{x})/T$$

and the *sample autocorrelation function* is given by $\widehat{\rho}_s = \widehat{\gamma}_s/\widehat{\gamma}_o$. Figure 14.2 plots $\widehat{\rho}_s$ against s for the consumption series. This is called the *sample correlogram*. For a stationary process, ρ_s declines sharply as the number of lags s increase. This is not necessarily the case for a nonstationary series. In the next section, we briefly review a popular method for the analysis of time-series known as the Box and Jenkins (1970) technique. This method utilizes the sample autocorrelation function to establish whether a series is stationary or not.

Sample: 1950 1993
Included observations: 44

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
. |*******	. |*******	1	0.941	0.941	41.683	0.000
. |*******	. |.	2	0.880	–0.045	79.040	0.000
. |******	. |.	3	0.820	–0.035	112.19	0.000
. |******	.* |.	4	0.753	–0.083	140.88	0.000
. |*****	.* |.	5	0.683	–0.064	165.11	0.000
. |*****	. |.	6	0.614	–0.034	185.21	0.000
. |****	. |.	7	0.547	–0.025	201.59	0.000
. |****	. |.	8	0.479	–0.049	214.50	0.000
. |***	. |.	9	0.412	–0.043	224.30	0.000
. |***	. |.	10	0.348	–0.011	231.52	0.000
. |**	. |.	11	0.288	–0.022	236.60	0.000
. |**	. |.	12	0.230	–0.025	239.94	0.000
. |*	.* |.	13	0.171	–0.061	241.85	0.000

Figure 14.2 Correlogram of Consumption

14.3 The Box and Jenkins Method

This method fits *Autoregressive Integrated Moving Average* (ARIMA) type models. We have already considered simple AR and MA type models in Chapters 5 and 6. The Box-Jenkins methodology differences the series and looks at the sample correlogram of the differenced series

to see whether stationarity is achieved. As will be clear shortly, if we have to difference the series once, twice or three times to make it stationary, this series is *integrated* of order 1, 2 or 3, respectively. Next, the Box-Jenkins method looks at the *autocorrelation function* and the *partial autocorrelation function* (synonymous with partial correlation coefficients) of the resulting stationary series to identify the order of the AR and MA process that is needed. The partial correlation between y_t and y_{t-s} is the correlation between those two variables holding constant the effect of all intervening lags, see Box and Jenkins (1970) for details. Figure 14.3 plots an AR(1) process of size $T = 250$ generated as $y_t = 0.7y_{t-1} + \epsilon_t$ with $\epsilon_t \sim \text{IIN}(0, 4)$. Figure 14.4 shows that the correlogram of this AR(1) process declines geometrically as s increases. Similarly, Figure 14.5 plots an MA(1) process of size $T = 250$ generated as $y_t = \epsilon_t + 0.4\epsilon_{t-1}$ with $\epsilon_t \sim \text{IIN}(0, 4)$. Figure 14.6 shows that the correlogram of this MA(1) process is zero after the first lag, see also problems 1 and 2 for further analysis. Identifying the right ARIMA model is not an exact science, but potential candidates emerge. These models are estimated using maximum likelihood techniques. Next, these models are subjected to some diagnostic checks. One commonly used check is to see whether the residuals are White noise. If they fail this test, these models are dropped from the list of viable candidates.

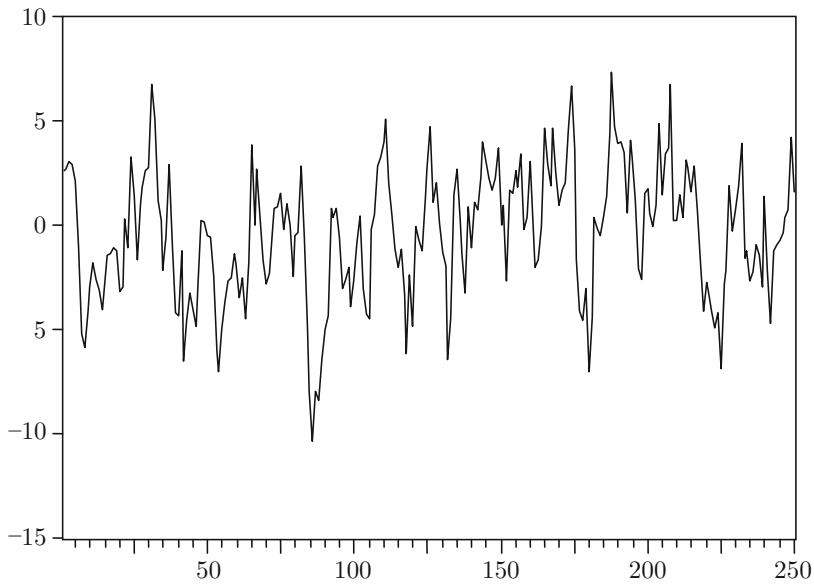


Figure 14.3 AR(1) Process, $\rho = 0.7$

If the time-series is *White noise*, i.e., purely random with constant mean and variance and zero autocorrelation, then $\rho_s = 0$ for $s > 0$. In fact, for a White noise series, if $T \to \infty$, $\sqrt{T}\widehat{\rho}_s$ will be asymptotically distributed $N(0, 1)$. A joint test for H_o; $\rho_s = 0$ for $s = 1, 2, \ldots, m$ lags, is given by the Box and Pierce (1970) statistic

$$Q = T \sum_{s=1}^{m} \widehat{\rho}_s^2 \tag{14.1}$$

This is asymptotically distributed under the null as χ^2_m. A refinement of the Box-Pierce Q-statistic that performs better, i.e., have more power in small samples is the Ljung and Box (1978) Q_{LB} statistic

Sample: 1 250
Included observations: 250

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
. |******	. |******	1	0.725	0.725	132.99	0.000
. |****	. |.	2	0.503	–0.048	197.27	0.000
. |***	. |.	3	0.330	–0.037	225.05	0.000
. |**	. |.	4	0.206	–0.016	235.92	0.000
. |*	. |.	5	0.115	–0.022	239.33	0.000
. |.	. |.	6	0.036	–0.050	239.67	0.000
. |.	. |.	7	–0.007	0.004	239.68	0.000
. |.	. |.	8	–0.003	0.050	239.68	0.000
. |.	. |.	9	–0.017	–0.041	239.75	0.000
* |.	* |.	10	–0.060	–0.083	240.71	0.000
* |.	* |.	11	–0.110	–0.063	243.91	0.000
. |.	. |*	12	–0.040	0.191	244.32	0.000

Figure 14.4 Correlogram of AR(1)

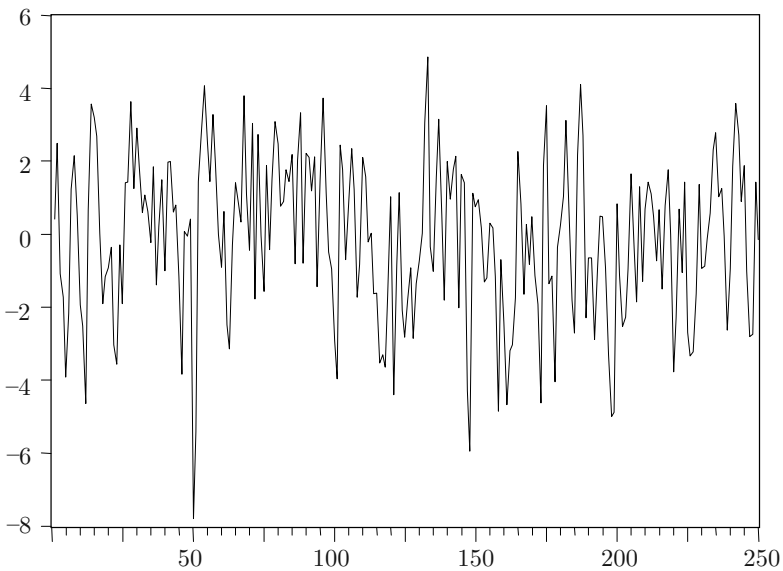


Figure 14.5 MA(1) Process, $\theta = 0.4$

$$Q_{LB} = T(T+2)\sum_{j=1}^{m} \widehat{\rho}_j^2/(T-j) \tag{14.2}$$

This is also distributed asymptotically as χ_m^2 under the null hypothesis. Maddala (1992, p. 540) warns about the inappropriateness of the Q and Q_{LB} statistics for autoregressive models. The arguments against their use are the same as those for not using the Durbin-Watson statistic in autoregressive models. Maddala (1992) suggests the use of LM statistics of the type proposed by Godfrey (1979) to evaluate the adequacies of the ARMA model proposed.

For the consumption series, $T = 42$ and the 95% confidence interval for $\widehat{\rho}_s$ is $0 \pm 1.96\ (1/\sqrt{42})$ which is ± 0.3024. Figure 14.2 plots this 95% confidence interval as two solid lines around zero.

Sample: 1 250
Included observations: 250

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
. |***	. |***	1	0.399	0.399	40.240	0.000
. |.	*|.	2	0.033	–0.150	40.520	0.000
. |.	. |*	3	0.010	0.066	40.545	0.000
. |.	. |.	4	–0.002	–0.033	40.547	0.000
. |*	. |*	5	0.090	0.127	42.624	0.000
. |.	. |.	6	0.055	–0.045	43.417	0.000
. |.	. |.	7	0.028	0.042	43.625	0.000
. |.	*|.	8	–0.031	–0.075	43.881	0.000
. |.	. |.	9	–0.034	0.023	44.190	0.000
. |.	. |.	10	–0.027	–0.045	44.374	0.000
. |.	. |.	11	–0.013	0.020	44.421	0.000
. |*	. |*	12	0.082	0.086	46.190	0.000

Figure 14.6 Correlogram of MA(1)

It is clear that $\widehat{\rho}_1$ up to $\widehat{\rho}_{10}$ are significantly different from zero and the sample correlogram declines slowly as the number of lags s increase. Moreover, the Q_{LB} statistics which are reported for lags 1, 2, up to 13 are all statistically significant. These were computed using EViews. Based on the sample correlogram and the Ljung-Box statistic, the consumption series is not purely random white noise. Figure 14.7 plots the sample correlogram for $\Delta C_t = C_t - C_{t-1}$. Note that this sample correlogram dies out abruptly after the first lag. Also, the Q_{LB} statistics are not significant after the first lag. This indicates stationarity of the first-differenced consumption series. Problem 3 asks the reader to plot the sample correlogram for personal disposable income and its first difference, and to compute the Ljung-Box Q_{LB} statistic to test for purely White noise based on 13 lags.

Sample: 1950 1993
Included observations: 43

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
. |***	. |***	1	0.344	0.344	5.4423	0.020
.* |.	.** |.	2	–0.067	–0.209	5.6518	0.059
.* |.	.* |.	3	–0.156	–0.066	6.8244	0.078
.* |.	. |.	4	–0.105	–0.038	7.3726	0.117
.* |.	.* |.	5	–0.077	–0.065	7.6740	0.175
.* |.	.* |.	6	–0.072	–0.060	7.9469	0.242
. |.	. |.	7	0.026	0.056	7.9820	0.334
. |.	.* |.	8	–0.050	–0.133	8.1177	0.422
. |.	. |.	9	0.058	0.133	8.3079	0.503
. |* .	. |.	10	0.073	–0.014	8.6244	0.568
. |* .	. |.	11	0.078	0.058	8.9957	0.622
. |.	.* |.	12	–0.033	–0.079	9.0637	0.697
. |.	. |.	13	–0.069	0.005	9.3724	0.744

Figure 14.7 Correlogram of First Difference of Consumption

A difficult question when modeling economic behavior is to decide on what lags should be in the ARIMA model, or the dynamic regression model. Granger et al. (1995) argue that there are disadvantages in using hypothesis testing to help make model specification decisions based on the data. They recommend instead the use of model selection criteria to make those decisions.

The Box-Jenkins methodology has been popular primarily among forecasters who claimed better performance than simultaneous equations models based upon economic theory. Box-Jenkins models are general enough to allow for nonstationarity and can handle seasonality. However, the Box-Jenkins models suffer from the fact that they are devoid of economic theory and as such they are not designed to test economic hypothesis, or provide estimates of key elasticity parameters. As a consequence, this method cannot be used for simulating the effects of a tax change or a Federal Reserve policy change. One lesson that economists learned from the Box-Jenkins methodology is that they have to take a hard look at the time-series properties of their variables and properly specify the dynamics of their economic models. Another popular forecasting technique in economics is the *Vector Autoregression* (VAR) methodology proposed by Sims (1980). This will be briefly discussed next.

14.4 Vector Autoregression

Sims (1980) criticized the simultaneous equation literature for the ad hoc restrictions needed for identification and for the ad hoc classification of exogenous and endogenous variables in the system, see Chapter 11. Instead, Sims (1980) suggested *Vector Autoregression* (VAR) models for forecasting macro time-series. VAR assumes that all the variables are endogenous. For example, consider the following three macro-series: money supply, interest rate, and output. VAR models this vector of three endogenous variables as an autoregressive function of their lagged values. VAR models can include some exogenous variables like trends and seasonal dummies, but the whole point is that it does not have to classify variables as endogenous or exogenous. If we allow 5 lags on each endogenous variable, each equation will have 16 parameters to estimate if we include a constant. For example, the money supply equation will be a function of 5 lags on money, 5 lags on the interest rate and 5 lags on output. Since the parameters are different for each equation the total number of parameters in this unrestricted VAR is $3 \times 16 = 48$ parameters. This degrees of freedom problem becomes more serious as the number of lags m and number of equations g increase. In fact, the number of parameters to be estimated becomes $g + mg^2$. With small samples, individual parameters may not be estimated precisely. So, only simple VAR models, can be considered for a short sample. Since this system of equations has the same set of variables in each equation SUR on the system is equivalent to OLS on each equation, see Chapter 10. Under normality of the disturbances, MLE as well as Likelihood Ratio tests can be performed. One important application of LR tests in the context of VAR is its use in determining the choice of lags to be used. In this case, one obtains the log-likelihood for the restricted model with m lags and the unrestricted model with $q > m$ lags. This LR test will be asymptotically distributed as $\chi^2_{(q-m)g^2}$. Once again, the sample size T should be large enough to estimate the large number of parameters $(qg^2 + g)$ for the unrestricted model.

One can of course impose restrictions to reduce the number of parameters to be estimated, but this reintroduces the problem of ad hoc restrictions which VAR was supposed to cure in the first place. Bayesian VAR procedures claim success with forecasting, see Litterman (1986), but again these models have been criticized because they are devoid of economic theory.

VAR models have also been used to test the hypothesis that some variables do not *Granger cause* some other variables.[1] For a two-equation VAR, as long as this VAR is correctly specified and no variables are omitted, one can test, for example, that y_1 does not Granger cause y_2. This hypothesis cannot be rejected if all the m lagged values of y_1 are insignificant in the equation for y_2. This is a simple F-test for the joint significance of the lagged coefficients of y_1 in the y_2 equation. This is asymptotically distributed as $F_{m,T-(2m+1)}$. The problem with the Granger test for non-causality is that it may be sensitive to the number of lags m, see Gujarati (1995). For an extensive analysis of nonstationary VAR models as well as testing and estimation of cointegrating relationships in VAR models, see Hamilton (1994) and Lütkepohl (2001).

14.5 Unit Roots

If $x_t = x_{t-1} + u_t$ where u_t is IID$(0, \sigma^2)$, then x_t is a *random walk*. Some stock market analysts believe that stock prices follow a random walk, i.e., the price of a stock today is equal to its price yesterday plus a random shock. This is a nonstationary time-series. Any shock to the price of this stock is *permanent* and does not die out like an AR(1) process. In fact, if the initial price of the stock is $x_o = \mu$, then

$$x_1 = \mu + u_1,\ x_2 = \mu + u_1 + u_2, \ldots, \quad \text{and} \quad x_t = \mu + \textstyle\sum_{j=1}^{t} u_j$$

with $E(x_t) = \mu$ and $\text{var}(x_t) = t\sigma^2$ since $u \sim \text{IID}(0, \sigma^2)$. Therefore, the variance of x_t is *dependent* on t and x_t is *not* covariance-stationary. In fact, as $t \to \infty$, so does $\text{var}(x_t)$. However, first differencing x_t we get u_t which is stationary. Figure 14.8 plots the graph of a random walk of size $T = 250$ generated as $x_t = x_{t-1} + \epsilon_t$ with $\epsilon_t \sim \text{IIN}(0, 4)$. Figure 14.9 shows that the autocorrelation function of this random walk process is persistent as s increases. Note that a random walk is an AR(1) model $x_t = \rho x_{t-1} + u_t$ with $\rho = 1$. Therefore, a test for nonstationarity is a test for $\rho = 1$ or a test for a *unit root*.

Using the lag operator L we can write the random walk as $(1 - L)x_t = u_t$ and in general, any autoregressive model in x_t can be written as $A(L)x_t = u_t$ where $A(L)$ is a polynomial in L. If $A(L)$ has $(1 - L)$ as one of its roots, then x_t has a unit root.

Subtracting x_{t-1} from both sides of the AR(1) model we get

$$\Delta x_t = (\rho - 1)x_{t-1} + u_t = \delta x_{t-1} + u_t \tag{14.3}$$

where $\delta = \rho - 1$ and $\Delta x_t = x_t - x_{t-1}$ is the first-difference of x_t. A test for H_o; $\rho = 1$ can be obtained by regressing Δx_t on x_{t-1} and testing that H_o; $\delta = 0$. Since u_t is stationary then if $\delta = 0$, $\Delta x_t = u_t$ and x_t is difference stationary, i.e., it becomes stationary after differencing it once. In this case, the original undifferenced series x_t is said to be integrated of order 1 or $I(1)$. If we need to difference x_t twice to make it stationary, then x_t is $I(2)$. A stationary process is by definition $I(0)$. Dickey and Fuller (1979) showed that the usual regression t-statistic for H_o; $\delta = 0$ from (14.3) does *not* have a t-distribution under H_o. In fact, this t-statistic has a non-standard distribution, see Bierens (2001) for a simple derivation of these results. Dickey and Fuller tabulated the critical values of the t-statistic $= (\widehat{\rho} - 1)/s.e.(\widehat{\rho}) = \widehat{\delta}/s.e.(\widehat{\delta})$ using Monte Carlo experiments. These tables have been extended by MacKinnon (1991). If $|t|$ exceeds the critical values, we reject H_o that $\rho = 1$ which also means that we do not reject the hypothesis of stationarity of the time-series. Non-rejection of H_o; $\rho = 1$ means that we do not reject the

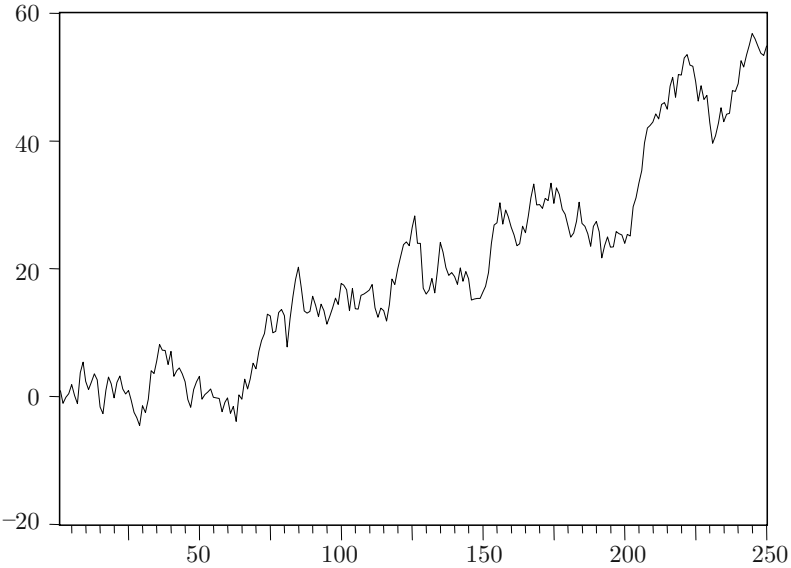


Figure 14.8 Random Walk Process

presence of a unit root and hence the nonstationarity of the time-series. Note that non-rejection of H_o may also be a non-rejection of $\rho = 0.99$. More formally stated, a weakness of unit root tests in general is that they have low power discriminating between a unit root process and a borderline stationary process. In practice, the *Dickey-Fuller* test has been applied in the following three forms:

$$\Delta x_t = \delta x_{t-1} + u_t \tag{14.4}$$

$$\Delta x_t = \alpha + \delta x_{t-1} + u_t \tag{14.5}$$

$$\Delta x_t = \alpha + \beta t + \delta x_{t-1} + u_t \tag{14.6}$$

where t is a time trend. The null hypothesis of the existence of a unit root is H_o; $\delta = 0$. This is the same for (14.4), (14.5) and (14.6), but the critical values for the corresponding t-statistics are different in each case. Standard time-series software like EViews give the proper critical values for the Dickey-Fuller statistic. For alternative unit root tests, see Phillips and Perron (1988) and Bierens and Guo (1993). In practice, one should run (14.6) if the series is trended with drift and (14.5) if it is trended without drift. Not including a constant or trend as in (14.4) is unlikely for economic data. The Box-Jenkins approach differences the series and looks at the sample correlogram of the differenced series. The Dickey-Fuller test is a more formal test for the existence of a unit root. Maddala (1992, p. 585) warns the reader to perform both the visual inspection and the unit root test before deciding on whether the time-series process is nonstationary.

If the disturbance term u_t follows a stationary AR(1) process, then the *augmented Dickey-Fuller* test runs the following modified version of (14.6) by including one additional regressor, Δx_{t-1}:

$$\Delta x_t = \alpha + \beta t + \delta x_{t-1} + \lambda \Delta x_{t-1} + \epsilon_t \tag{14.7}$$

Sample: 1 250
Included observations: 250

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
. |********	. |********	1	0.980	0.980	242.76	0.000
. |*******	. |.	2	0.959	–0.003	476.56	0.000
. |*******	. |.	3	0.940	0.004	701.83	0.000
. |*******	. |.	4	0.920	–0.013	918.61	0.000
. |*******	. |.	5	0.899	–0.044	1126.4	0.000
. |*******	. |.	6	0.876	–0.053	1324.6	0.000
. |*******	. |.	7	0.855	0.028	1514.2	0.000
. |******	. |*	8	0.837	0.067	1696.7	0.000
. |******	. |.	9	0.821	0.032	1872.8	0.000
. |******	. |.	10	0.804	–0.006	2042.6	0.000
. |******	. |.	11	0.788	–0.007	2206.3	0.000
. |******	. |.	12	0.774	0.030	2364.8	0.000

Figure 14.9 Correlogram of a Random Walk Process

In this case, the t-statistic for $\delta = 0$ is a unit root test allowing for first-order serial correlation. This augmented Dickey-Fuller test in (14.7) has the same asymptotic distribution as the corresponding Dickey-Fuller test in (14.6) and the same critical values can be used. Similarly, if u_t follows a stationary AR(p) process, this amounts to adding p extra regressors in (14.6) consisting of $\Delta x_{t-1}, \Delta x_{t-2}, \ldots, \Delta x_{t-p}$ and testing that the coefficient of x_{t-1} is zero. In practice, one does not know the process generating the serial correlation in u_t and the general practice is to include as many lags of Δx_t as is necessary to render the ϵ_t term in (14.7) serially uncorrelated. More lags may be needed if the disturbance term contains Moving Average terms, since a MA term can be thought of as an infinite autoregressive process, see Ng and Perron (1995) for an extensive Monte Carlo on the selection of the truncation lag. Two other important complications when doing unit root tests are: (i) structural breaks in the time-series, like the oil embargo of 1973, tend to bias the standard unit root tests against rejecting the null hypothesis of a unit root, see Perron (1989). (ii) Seasonally adjusted data also tend to bias the standard unit root tests against rejecting the null hypothesis of a unit root, see Ghysels and Perron (1992). For this reason, Davidson and MacKinnon (1993, p. 714) suggest using seasonally unadjusted data whenever available.

For the trended consumption series with drift, the following regression is performed:

$$\begin{underset{\quad (2.96)}{\Delta C_t = 1088.86 +} \; \underset{(2.79)}{39.90\; t} - \underset{(2.64)}{0.1956\; C_{t-1}} + \text{residuals} \end{underset} \tag{14.8}$$

where the numbers in parentheses are the usually reported t-statistics. The null hypothesis is that the coefficient of C_{t-1} in this regression is zero. Table 14.1 gives the Dickey-Fuller t-statistic (-2.64) and the corresponding 5% critical value (-3.5162) tabulated by MacKinnon (1991). This is done using EViews. Note that @TREND(1950) is the time-trend starting at 1950. Since the computed t-statistic for the coefficient of C_{t-1} does not exceed its corresponding critical value, we do not reject the null hypothesis of the existence of a unit root. We conclude that C_t is nonstationary. This confirms our finding from the sample correlogram of C_t given in Figure 14.2.

Table 14.1 Dickey-Fuller Test

ADF Test Statistic	–2.644192	1% Critical Value*	–4.1837
		5% Critical Value	–3.5162
		10% Critical Value	–3.1882

* MacKinnon critical values for rejection of hypothesis of a unit root.

Augmented Dickey-Fuller Test Equation	
Dependent Variable:	D(CONSUM)
Method:	Least Squares
Sample (adjusted):	1951 1993
Included observations:	43 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
CONSUM(–1)	–0.195606	0.073976	–2.644192	0.0116
C	1088.862	367.6877	2.961377	0.0051
@ TREND(1950)	39.90164	14.31344	2.787705	0.0081

R-squared	0.178423	Mean dependent var	176.0698
Adjusted R-squared	0.137344	S.D. dependent var	158.7145
S.E. of regression	147.4129	Akaike info criterion	12.89157
Sum squared resid	869223.0	Schwarz criterion	13.01444
Log likelihood	–274.1687	F-statistic	4.343415
Durbin-Watson stat	1.295028	Prob (F-statistic)	0.019632

Allowing for serial correlation in the disturbance term, we included ΔC_{t-1} and ΔC_{t-2} in (14.8) to get

$$\begin{underset{}{\Delta C_t =} \underset{(3.12)}{1251.84 +} \underset{(2.93)}{48.64\ t -} \underset{(2.86)}{0.241\ C_{t-1} +} \underset{(2.96)}{0.436\ \Delta C_{t-1} -} \underset{(0.63)}{0.097\ \Delta C_{t-2}} + \text{residuals}} \tag{14.9}$$

The coefficient of C_{t-1} has a t-statistic of -2.86 which still exceeds the 5% critical value of -3.5217. Hence, we cannot reject the hypothesis that C_t is nonstationary even after allowing for second-order serial correlation in the disturbances.

One can check whether the first-differenced series is stationary by performing a unit root test on the first-differenced model. Let $\widetilde{C}_t = \Delta C_t$, then run the following regression:

$$\begin{underset{}{\Delta \widetilde{C}_t =} \underset{(3.46)}{119.35 -} \underset{(4.43)}{0.652\ \widetilde{C}_{t-1}} + \text{residuals}} \tag{14.10}$$

the coefficient of $\widetilde{C}_{t-1}$ has a t-statistic of -4.43 which is smaller than the 5% critical value of -2.9320. In other words, we reject the hypothesis of unit root for the first-differenced series ΔC_t. This confirms our finding from the sample correlogram of ΔC_t given in Figure 14.7 and we conclude that C_t is $I(1)$.

So far, all tests for unit root have the hypothesis of nonstationarity as the null with the alternative being that the series is stationary. Two unit roots tests with stationarity as the null and nonstationarity as the alternative are given by Kwaitowski et al. (1992) and Leybourne and McCabe (1994). The first test known as KPSS is an analog of the Phillips-Perron test whereas the Leybourne-McCabe test is an analog of the augmented Dickey-Fuller test. Reversing the null may lead to confirmation of stationarity or nonstationarity or may yield conflicting decisions.

14.6 Trend Stationary versus Difference Stationary

Many macroeconomic time-series that are trending upwards have been characterized as either

$$\text{Trend Stationary: } x_t = \alpha + \beta t + u_t \tag{14.11}$$

or

$$\text{Difference Stationary: } x_t = \gamma + x_{t-1} + u_t \tag{14.12}$$

where u_t is stationary. The first model (14.11) says that the macro-series is stationary except for a *deterministic* trend. $E(x_t) = \alpha + \beta t$ which varies with t. In contrast, the second model (14.12) says that the macro-series is a *random walk with drift*. The drift parameter γ in (14.12) plays the same role as the β parameter in (14.11), since both cause x_t to trend upwards over time. Model (14.11) is consistent with economists introducing a time trend in the regression. This has the same effect as detrending each variable in the regression rendering it stationary, see the Frisch-Waugh-Lovell Theorem in Chapter 7. This detrending is valid only if model (14.11) is true for every series in the regression. Model (14.12) on the other hand, requires differencing to obtain a stationary series. Detrending and differencing are two completely different remedies. What is valid for one model is not valid for the other. The choice between (14.11) and (14.12) is based on a test for the existence of a unit root. Essential reading on these two models are Nelson and Plosser (1982) and Stock and Watson (1988). Nelson and Plosser applied the Dickey-Fuller test to a wide range of historical macro time-series for the U.S. economy and found that all of these series were difference stationary, with the exception of the unemployment rate. Plosser and Schwert (1978) argued that for most economic macro time-series, it is best to difference the data rather than work with levels. The reasoning is that if these series are difference stationary and we run regressions in levels, the usual properties of our estimators as well as the distributions of the associated test statistics are invalid. On the other hand, if the true model is a regression in levels with the data series being trend stationary, differencing the model will produce a Moving Average error term and at worst, ignoring it will lead to loss in efficiency. It is important to emphasize, that for nonstationary variables, the standard asymptotic theory does not apply, see problems 6 and 7, and that t and F-statistics obtained from regressions using these variables may have non-standard distributions, see Durlauf and Phillips (1988).

Granger and Newbold (1974) demonstrated some of the problems associated with regressing nonstationary time-series on each other. In fact, they showed that if x_t and y_t are independent random walks, then one should expect to find no evidence of a relationship when one regresses y_t on x_t. In other words, the estimate of β in the regression $y_t = \alpha + \beta x_t + u_t$ should be near zero and its associated t-statistic insignifiant. In fact, for a sample size of 50 and a 100 replications, Granger and Newbold found $|t| \leq 2$ on only 23 occasions, $2 < |t| \leq 4$ on 24 occasions, and $|t| > 4$ on 53 occasions. Granger and Newbold (1974) called this phenomenon *spurious regression* since it finds a significant relationship between the two time-series when none exists. Hence, one should be cautious when running time-series regressions involving unit root processes. High R^2 and significant t-statistics from OLS regressions may be hiding nonsense results. Phillips (1986) studied the asymptotic properties of the least squares spurious regression model and confirmed these simulation results. In fact, Phillips showed that the t-statistic for H_o; $\beta = 0$ converges in probability to ∞ as $T \to \infty$. This means that the t-statistic will reject H_o; $\beta = 0$ with probability 1 as $T \to \infty$. If both x_t and y_t are independent trend stationary series generated as

described in (14.11), then the R^2 of the regression of y_t on x_t will tend to one as $T \to \infty$, see Davidson and MacKinnon (1993, p. 671). For a summary of several extensions of these results, see Granger (2001).

14.7 Cointegration

Let us continue with our consumption-income example. In Chapter 5, we regressed C_t on Y_t and obtained

$$\begin{matrix} C_t = -65.8 + & 0.916 & Y_t + \text{residuals} \\ \quad (0.72) & \quad (105.9) & \end{matrix} \tag{14.13}$$

with $R^2 = 0.996$ and D.W. $= 0.46$. We have shown that C_t and Y_t are nonstationary series and that both are $I(1)$, see also problem 3. The regression in (14.13) could be a *spurious regression* owing to the fact that we regressed a nonstationary series on another. This invalidates the t and F-statistics of regression (14.13). Since both C_t and Y_t are integrated of the same order, and Figure 14.1 shows that they are trending upwards together, this random walk may be in unison. This is the idea behind *cointegration*. C_t and Y_t are cointegrated if there exists a linear combination of C_t and Y_t that yields a stationary series. More formally, if C_t and Y_t are both $I(1)$ but there exist a linear combination $C_t - \alpha - \beta Y_t = u_t$ which is $I(0)$, then C_t and Y_t are *cointegrated* and β is the cointegrating parameter. This idea can be extended to a vector of more than two time-series. This vector is cointegrated if the components of this vector have a unit root and there exists a linear combination of this vector that is stationary. Such a cointegrating relationship can be interpreted as a stable long-run relationship between the components of this time-series vector. Economic examples of long-run relationship include the quantity theory of money, purchasing power parity and the permanent income theory of consumption. The important point to emphasize here is that differencing these nonstationary time-series destroys potential valuable information about the long-run relationship between these economic variables. The theory of cointegration tries to estimate this long-run relationship using the nonstationary series themselves, rather than their first differences. In order to explain this, we state (without proof) one of the implications of the *Granger Representation Theorem*, namely, that a set of cointegrated variables will have an *Error-Correction Model* (ECM) representation. Let us illustrate with an example.

A Cointegration Example

This is based on Engle and Granger (1987). Assume that C_t and Y_t for $t = 1, 2, \ldots, T$ are $I(1)$ processes generated as follows:

$$C_t - \beta Y_t = u_t \quad \text{with} \quad u_t = \rho u_{t-1} + \epsilon_t \quad \text{and} \quad |\rho| < 1 \tag{14.14}$$

$$C_t - \alpha Y_t = \nu_t \quad \text{with} \quad \nu_t = \nu_{t-1} + \eta_t \quad \text{and} \quad \alpha \neq \beta \tag{14.15}$$

In other words, u_t follows a stationary AR(1) process, while ν_t follows a random walk. Suppose that $\begin{pmatrix} \epsilon_t \\ \eta_t \end{pmatrix}$ are independent bivariate normal random variables with mean zero and variance $\Sigma = [\sigma_{ij}]$ for $i, j = 1, 2$. First, we obtain the *reduced form* representation of Y_t and C_t in terms

of u_t and ν_t. This is given by

$$C_t = \frac{\alpha}{(\alpha-\beta)}u_t + \frac{\beta}{(\alpha-\beta)}\nu_t \tag{14.16}$$

$$Y_t = \frac{1}{(\alpha-\beta)}u_t - \frac{1}{(\alpha-\beta)}\nu_t \tag{14.17}$$

Since u_t is $I(0)$ and ν_t is $I(1)$, we conclude from (14.16) and (14.17) that C_t and Y_t are in fact $I(1)$ series. In terms of the usual *order condition* for identification considered in Chapter 11, the system of equations given by (14.14) and (14.15) are not identified because there are no exclusion restrictions on either equation. However, if we take a linear combination of the two structural equations given in (14.14) and (14.15), the disturbance of the resulting linear combination is neither a stationary AR(1) process nor a random walk. Hence, both (14.14) and (14.15) are identified. Note that if $\rho = 1$, then u_t is a random walk and the linear combination of u_t and ν_t is also a random walk. In this case, neither (14.14) nor (14.15) are identified.

In the Engle-Granger terminology, $C_t - \beta Y_t$ is the cointegrating relationship and $(1, -\beta)$ is the cointegrating vector. This cointegrating relationship is unique. The proof is by contradiction. Assume there is another cointegrating relationship $C_t - \gamma Y_t$ that is $I(0)$, then the difference between the two cointegrating relationships yields $(\gamma - \beta)Y_t$. This is also $I(0)$. This can only happen for every value of Y_t, which is $I(1)$, if and only if $\beta = \gamma$.

Difference both equations in (14.14) and (14.15) and write both differenced equations as a system of two equations in $(\Delta C_t, \Delta Y_t)'$, one gets:

$$\begin{bmatrix} 1 & -\beta \\ 1 & -\alpha \end{bmatrix} \begin{bmatrix} \Delta C_t \\ \Delta Y_t \end{bmatrix} = \begin{bmatrix} \Delta u_t \\ \Delta \nu_t \end{bmatrix} = \begin{bmatrix} \epsilon_t + (\rho-1)C_{t-1} - \beta(\rho-1)Y_{t-1} \\ \eta_t \end{bmatrix} \tag{14.18}$$

where the second equality is obtained by replacing $\Delta\nu_t$ by η_t, Δu_t by $(\rho-1)u_{t-1} + \epsilon_t$, and substituting for u_{t-1} its value $(C_{t-1} - \beta Y_{t-1})$. Post-multiplying (14.18) by the inverse of the first matrix, one can show, see problem 9, that the resulting solution is the following VAR model:

$$\begin{bmatrix} \Delta C_t \\ \Delta Y_t \end{bmatrix} = \frac{1}{(\beta-\alpha)} \begin{bmatrix} -\alpha(\rho-1) & \alpha\beta(\rho-1) \\ -(\rho-1) & \beta(\rho-1) \end{bmatrix} \begin{pmatrix} C_{t-1} \\ Y_{t-1} \end{pmatrix} + \begin{pmatrix} h_t \\ g_t \end{pmatrix} \tag{14.19}$$

where h_t and g_t are linear combinations of ϵ_t and η_t. Note that if $\rho = 1$, then the level variables C_{t-1} and Y_{t-1} drop from the VAR equations. Let $Z_t = C_t - \beta Y_t$ and define $\delta = (\rho-1)/(\beta-\alpha)$. Then the VAR representation in (14.19) can be written as follows:

$$\Delta C_t = -\alpha\delta Z_{t-1} + h_t \tag{14.20}$$

$$\Delta Y_t = -\delta Z_{t-1} + g_t \tag{14.21}$$

This is the *Error-Correction Model* (ECM) representation of the original model. Z_{t-1} is the error correction term. It represents a disequilibrium term showing the departure from long-run equilibrium, see section 6.4. Note that if $\rho = 1$, then $\delta = 0$ and Z_{t-1} drops from both ECM equations. As Banerjee et al. (1993, p. 139) explain, this ECM representation is a noteworthy "...contribution to resolving, or synthesizing, the debate between time-series analysts and those favoring econometric methods." The former considered only differenced time-series that can be legitimately assumed stationary, while the latter focused on equilibrium relationships expressed

in levels. The former wiped out important long-run relationships by first differencing them, while the latter ignored the spurious regression problem. In contrast, the ECM allows the use of first differences and levels from the cointegrating relationship. For more details, see Banerjee et al. (1993). A simple two-step procedure for estimating cointegrating relationships is given by Engle and Granger (1987). In the first step, the OLS estimator of β is obtained by regressing C_t on Y_t. This can be shown to be *superconsistent*, i.e., plim $T(\widehat{\beta}_{OLS} - \beta) \to 0$ as $T \to \infty$. Using $\widehat{\beta}_{OLS}$ one obtains $\widehat{Z}_t = C_t - \widehat{\beta}_{OLS} Y_t$. In the second step, using $\widehat{Z}_{t-1}$ rather than Z_{t-1}, apply OLS to estimate the ECM in (14.20) and (14.21). Extensive Monte Carlo experiments have been conducted by Banerjee et al. (1993) to investigate the bias of β in small samples. This is pursued further in problem 9. An alternative estimation procedure is the maximum likelihood approach suggested by Johansen (1988). This is beyond the scope of this book. See Dolado et al. (2001) for a lucid summary of the cointegration literature.

A formal test for cointegration is given by Engle and Granger (1987) who suggest running regression (14.13) and testing that the residuals do not have a unit root. In other words, run a Dickey-Fuller test or its augmented version on the resulting residuals from (14.13). In fact, if C_t and Y_t are not cointegrated, then any linear combination of them would be nonstationary including the residuals of (14.13). Since these tests are based on residuals, their asymptotic distributions are not the same as those of the corresponding ordinary unit roots tests. Asymptotic critical values for these tests can be found in Davidson and MacKinnon (1993, p. 722). For our consumption regression the following Dickey-Fuller test is obtained on the residuals:

$$\begin{aligned} \Delta\widehat{u}_t &= \underset{}{2.940} - \underset{}{0.207}\ \widehat{u}_t + \text{residuals} \\ &\quad\ (0.19) \quad (1.92) \end{aligned} \tag{14.22}$$

the Davidson and MacKinnon (1993) asymptotic 5% critical value for this t-statistic is -3.34. Since the observed t-value is larger than the critical value, we cannot reject the hypothesis that $\widehat{u}_t$ is nonstationary. We have also included a trend and two lags of the first-differenced residuals. All of the resulting augmented Dickey-Fuller tests did not reject the existence of a unit root. Therefore, C_t and Y_t are *not* cointegrated. This suggests that the relationship estimated in (14.13) is spurious. Regressing an $I(1)$ series on another lead to spurious results unless they are cointegrated. Of course, other $I(1)$ series may have been erroneously excluded from (14.13) which when included may result in a cointegrating relationship among the resulting variables. In other words, C_t and Y_t may *not* be cointegrated because of an omitted variables problem.

14.8 Autoregressive Conditional Heteroskedasticity

Financial time-series such as foreign exchange rates, inflation rates and stock prices may exhibit some volatility which varies over time. In the case of inflation or foreign exchange rates this could be due to changes in the Federal Reserve's policies. In the case of stock prices this could be due to rumors about a certain company's merger or takeover. This suggests that the variance of these time-series may be heteroskedastic. Engle (1982) modeled this heteroskedasticity by relating the conditional variance of the disturbance term at time t to the size of the squared disturbance terms in the recent past. A simple *Autoregressive Conditionally Heteroskedastic* (ARCH) model is given by

$$\sigma_t^2 = E(u_t^2/\zeta_t) = \gamma_o + \gamma_1 u_{t-1}^2 + .. + \gamma_p u_{t-p}^2 \tag{14.23}$$

where ζ_t denotes the information set upon which the variance of u_t is to be conditioned. This typically includes all the information available prior to period t. In (14.23), the variance of u_t conditional on the information prior to period t is an autoregressive function of order p in squared lagged values of u_t. This is called an ARCH(p) process. Since (14.23) is a variance, this means that all the γ_i's for $i = 0, 1, \ldots, p$ have to be non-negative. Engle (1982) showed that a simple test for homoskedasticity, i.e., H_o; $\gamma_1 = \gamma_2 = .. = \gamma_p = 0$, can be based upon an ordinary F-test which regresses the squared OLS residuals (e_t^2) on their lagged values ($e_{t-1}^2, \ldots, e_{t-p}^2$) and a constant. The F-statistic tests the joint significance of the regressors and is reported by most regression packages. Alternatively, one can compute T times the centered R^2 of this regression and this is distributed as χ_p^2 under the null hypothesis H_o. This test resembles the usual homoskedasticity tests studied in Chapter 5 except that the squared OLS residuals are regressed upon their lagged values rather than some explanatory variables.

The simple ARCH(1) process

$$\sigma_t^2 = \gamma_o + \gamma_1 u_{t-1}^2 \tag{14.24}$$

can be generated as follows: $u_t = [\gamma_o + \gamma_1 u_{t-1}^2]^{1/2} \epsilon_t$ where $\epsilon_t \sim \text{IID}(0, 1)$. Note that the simplifying variance of unity for ϵ_t can be achieved by rescaling the parameters γ_o and γ_1. In this case, the conditional mean of u_t is given by

$$E(u_t/\zeta_t) = [\gamma_o + \gamma_1 u_{t-1}^2]^{1/2} E(\epsilon_t/\zeta_t) = 0$$

since u_{t-1}^2 is known at time t. Similarly, the conditional variance can be easily obtained from

$$E(u_t^2/\zeta_t) = [\gamma_o + \gamma_1 u_{t-1}^2] E(\epsilon_t^2/\zeta_t) = \gamma_o + \gamma_1 u_{t-1}^2$$

since $E(\epsilon_t^2) = 1$. Also, the conditional covariances can be easily shown to be zero since

$$E(u_t u_{t-s}/\zeta_t) = u_{t-s} E(u_t/\zeta_t) = 0 \quad \text{for} \quad s = 1, 2, \ldots, t.$$

The unconditional mean can be obtained by taking repeated conditional expectations period by period until we reach the initial period, see the Appendix to Chapter 2. For example, taking the conditional expectation of $E(u_t/\zeta_t)$ based on information prior to period $t-1$, we get

$$E[E(u_t/\zeta_t)/\zeta_{t-1}] = E(0/\zeta_{t-1}) = 0$$

It is clear that all prior conditional expectations of zero will be zero so that $E(u_t) = 0$. Similarly, taking the conditional expectations of $E(u_t^2/\zeta_t)$ based on information prior to period $t-1$, we get

$$E[E(u_t^2/\zeta_t)/\zeta_{t-1}] = \gamma_o + \gamma_1 E[u_{t-1}^2/\zeta_{t-1}] = \gamma_o + \gamma_1(\gamma_o + \gamma_1 u_{t-2}^2) = \gamma_o(1 + \gamma_1) + \gamma_1^2 u_{t-2}^2$$

By taking repeated conditional expectations one period at a time we finally get

$$E(u_t^2) = \gamma_o(1 + \gamma_1 + \gamma_1^2 + .. + \gamma_1^{t-1}) + \gamma_1^t u_o^2$$

As $t \to \infty$, the unconditional variance of u_t is given by $\sigma^2 = \text{var}(u_t) = \gamma_o/(1 - \gamma_1)$ for $|\gamma_1| < 1$ and $\gamma_o > 0$. Therefore, the ARCH(1) process is homoskedastic.

ARCH models can be estimated using feasible GLS or maximum likelihood methods. Alternatively, one can use a double-length regression procedure suggested by Davidson and MacKinnon (1993) to obtain (i) one-step efficient estimates starting from OLS estimates or (ii) the maximum likelihood estimates. Here we focus on the feasible GLS procedure suggested by Engle (1982). For the regression model

$$y = X\beta + u \tag{14.25}$$

where y is $T \times 1$ and X is $T \times k$. First, obtain the OLS estimates $\widehat{\beta}_{OLS}$ and the OLS residuals e. Second, perform the following regression: $e_t^2 = a_o + a_1 e_{t-1}^2 +$ residuals. This yields a test for homoskedasticity. Third, compute $\widehat{\sigma}_t^2 = a_o + a_1 e_{t-1}^2$ and regress $[(e_t^2/\widehat{\sigma}_t) - 1]$ on $(1/\widehat{\sigma}_t)$ and $(e_{t-1}^2/\widehat{\sigma}_t)$. Call the regression estimates d_a. One updates $a' = (a_o, a_1)$ by computing $\widehat{a} = a + d_a$. Fourth, recompute $\widehat{\sigma}_t^2$ using the updated $\widehat{a}$ from step 3, and form the set of regressors $x_{tj} r_t$ for $j = 1, \ldots, k$, where

$$r_t = \left[\frac{1}{\widehat{\sigma}_t} + 2 \left(\frac{\widehat{a}_1 e_t}{\widehat{\sigma}_{t+1}} \right)^2 \right]^{1/2} \tag{14.26}$$

Finally, regress $(e_t s_t / r_t)$ where

$$s_t = \frac{1}{\widehat{\sigma}_t} - \frac{\widehat{a}_1}{\widehat{\sigma}_{t+1}} \left(\frac{e_{t+1}^2}{\widehat{\sigma}_{t+1}} - 1 \right)$$

on $x_{tj} r_t$ for $j = 1, \ldots, k$ and obtain the least squares coefficients d_β. Update the estimate of β by computing $\widehat{\beta} = \widehat{\beta}_{OLS} + d_\beta$. This procedure can run into problems if the $\widehat{\sigma}_t^2$ are not all positive, see Judge et al. (1985) and Engle (1982) for details.

The ARCH model has been generalized by Bollerslev (1986). The Generalized ARCH (GARCH (p, q)) model can be written as

$$\sigma_t^2 = \gamma_o + \textstyle\sum_{i=1}^{p} \gamma_i u_{t-i}^2 + \sum_{j=1}^{q} \delta_j \sigma_{t-j}^2 \tag{14.27}$$

In this case, the conditional variance of u_t depends upon q of its lagged values as well as p squared lagged values of u_t. The simple GARCH $(1, 1)$ model is given by

$$\sigma_t^2 = \gamma_o + \gamma_1 u_{t-1}^2 + \delta_1 \sigma_{t-1}^2 \tag{14.28}$$

An LM test for GARCH (p, q) turns out to be equivalent to testing ARCH $(p + q)$. This simply regresses squared OLS residuals on $(p + q)$ of its squared lagged values. The test statistic is T times the uncentered R^2 and is asymptotically distributed as χ_{p+q}^2 under the null of homoskedasticity.

For the consumption-income data, we regressed the squared residuals from (14.13) on their lagged values to test for homoskedasticity assuming an ARCH(1) model. We obtained

$$\begin{aligned} e_t^2 = {} & 13012.14 + \ 0.47 \ e_{t-1}^2 + \text{residuals} \\ & (2.47) \qquad (2.96) \end{aligned} \tag{14.29}$$

The observed F-statistic for this regression is 8.77. This is distributed as $F_{1,41}$ under the null hypothesis of homoskedasticity. The p-value for this observed F is 0.005. The LM statistic for

Table 14.2 ARCH Test

F-statistic	8.774600	Probability	0.005064	
Obs*R-squared	7.580328	Probability	0.005901	
Test Equation:				
Dependent Variable:	RESIDˆ2			
Method:	Least Squares			
Sample(adjusted):	1951 1993			
Included observations:	43 after adjusting endpoints			
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13012.14	5260.785	2.473422	0.0176
RESIDˆ2(–1)	0.467898	0.157956	2.962195	0.0051
R-squared	0.176287	Mean dependent var		22639.39
Adjusted R-squared	0.156196	S.D. dependent var		29530.97
S.E. of regression	27126.79	Akaike info criterion		23.29983
Sum squared resid	3.02E+10	Schwarz criterion		23.38174
Log likelihood	–498.9463	F-statistic		8.774600
Durbin-Watson stat	1.709016	Prob (F-statistic)		0.005064

the same hypothesis, obtained as T times the uncentered R^2, is 7.58. This is asymptotically distributed as χ^2_1 under the null hypothesis and has a p-value of 0.006. Hence, we reject the null of homoskedasticity. This can be done using EViews by clicking on residual tests for the regression in (14.13) and choosing the ARCH LM test using one lag as an option. This is shown in Table 14.2.

In conclusion, a lot of basic concepts have been introduced in this chapter and we barely scratched the surface. Hopefully, this will motivate the reader to take the next econometrics time series course.

Note

1. Granger causality has been developed by Granger (1969). For another definition of causality, see Sims (1972). Also, Chamberlain (1982) for a discussion on when these two definitions are equivalent.

Problems

1. For the AR(1) model

$$y_t = \rho y_{t-1} + \epsilon_t \quad t = 1, 2, \ldots, T; \quad \text{with} \quad |\rho| < 1 \quad \text{and} \quad \epsilon_t \sim \text{IIN}(0, \sigma^2_\epsilon)$$

   (a) Show that if $y_o \sim N(0, \sigma^2_\epsilon/1-\rho^2)$, then $E(y_t) = 0$ for all t and $\text{var}(y_t) = \sigma^2_\epsilon/(1-\rho^2)$ so that the mean and variance are independent of t. Note that if $\rho = 1$ then $\text{var}(y_t)$ is ∞. If $|\rho| > 1$ then $\text{var}(y_t)$ is negative!

   (b) Show that $\text{cov}(y_t, y_{t-s}) = \rho^s \sigma^2$ which is only dependent on s, the distance between the two time periods. Conclude from parts (a) and (b) that this AR(1) model is *weakly stationary.*

(c) Generate the above AR(1) series for $T = 250$, $\sigma_\epsilon^2 = 0.25$ and various values of $\rho = \pm 0.9, \pm 0.8, \pm 0.5, \pm 0.3$ and ± 0.1. Plot the AR(1) series and the autocorrelation function ρ_s versus s.

2. For the MA(1) model

$$y_t = \epsilon_t + \theta\epsilon_{t-1} \quad t = 1, 2, \ldots, T; \quad \text{with} \quad \epsilon_t \sim \text{IIN}(0, \sigma_\epsilon^2)$$

(a) Show that $E(y_t) = 0$ and $\text{var}(y_t) = \sigma_\epsilon^2(1+\theta^2)$ so that the mean and variance are independent of t.

(b) Show that $\text{cov}(y_t, y_{t-1}) = \theta\sigma_\epsilon^2$ and $\text{cov}(y_t, y_{t-s}) = 0$ for $s > 1$ which is only dependent on s, the distance between the two time periods. Conclude from parts (a) and (b) that this MA(1) model is *weakly stationary*.

(c) Generate the above MA(1) series for $T = 250$, $\sigma_\epsilon^2 = 0.25$ and various values of $\theta = 0.9$, 0.8, 0.5, 0.3 and 0.1. Plot the MA(1) series and the autocorrelation function versus s.

3. Using the consumption-personal disposable income data for the U.S. used in this chapter:

(a) Compute the sample autocorrelation function for personal disposable income (Y_t) for $m = 13$ lags. Plot the sample correlogram. Repeat for the first-differenced series (ΔY_t).

(b) Using a Ljung-Box Q_{LB} statistic, test that H_o; $\rho_s = 0$ for $s = 1, \ldots, 13$.

(c) Run the Dickey-Fuller regression given in (14.6) and test for the existence of a unit root in personal disposable income (Y_t).

(d) Run the augmented Dickey-Fuller regression in (14.7) adding one lag, two lags and three lags of ΔY_t to the right hand side of the regression.

(e) Define $\widetilde{Y}_t = \Delta Y_t$ and run $\Delta\widetilde{Y}_t$ on $\widetilde{Y}_{t-1}$ and a constant. Test that the first-differenced series of personal disposable income is stationary. What do you conclude? Is Y_t an $I(1)$ process?

(f) Replicate the regression in (14.22) and verify the Engle-Granger (1987) test for cointegration.

(g) Test for homoskedasticity assuming an ARCH(2) model for the disturbances of (14.13).

(h) Repeat parts (a) through (g) using log$C$ and log$Y$. Are there any changes in the above results?

4. (a) Generate $T = 25$ observations on x_t and y_t as independent random walks with IIN$(0, 1)$ disturbances. Run the regression $y_t = \alpha + \beta x_t + u_t$ and test the null hypothesis H_o; $\beta = 0$ using the usual t-statistic at the 1%, 5% and 10% levels. Repeat this experiment 1000 times and report the frequency of rejections at each significance level. What do you conclude?

(b) Repeat part (a) for $T = 100$ and $T = 500$.

(c) Repeat parts (a) and (b) generating x_t and y_t as independent random walks with drift as described in (14.13), using IIN$(0, 1)$ disturbances. Let $\gamma = 0.2$ for both series.

(d) Repeat parts (a) and (b) generating x_t and y_t as independent trend stationary series as described in (14.11), using IIN$(0, 1)$ disturbances. Let $\alpha = 1$ and $\beta = 0.04$ for both series.

(e) Report the frequency distributions of the R^2 statistics obtained in parts (a) through (d) for each sample size and method of generating the time-series. What do you conclude? **Hint:** See the Monte Carlo experiments in Granger and Newbold (1974), Davidson and MacKinnon (1993) and Banerjee, Dolado, Galbraith and Hendry (1993).

5. For the Money Supply, GNP and interest rate series data for the U.S. given on the Springer web site as MACRO.ASC, fit a VAR three equation model using:

(a) Two lags on each variable.

(b) Three lags on each variable.

(c) Compute the Likelihood Ratio test for part (a) versus part (b).

(d) For the two-equation VAR of Money Supply and interest rate with three lags on each variable, test that the interest rate does not Granger cause the money supply?

(e) How sensitive are the tests in part (d) if we had used only two lags on each variable.

6. For the *simple Deterministic Time Trend Model*

$$y_t = \alpha + \beta t + u_t \qquad t = 1, .., T$$

where $u_t \sim \text{IIN}(0, \sigma^2)$.

(a) Show that

$$\begin{pmatrix} \widehat{\alpha}_{OLS} - \alpha \\ \widehat{\beta}_{OLS} - \beta \end{pmatrix} = (X'X)X'u = \begin{bmatrix} T & \sum_{t=1}^{T} t \\ \sum_{t=1}^{T} t & \sum_{t=1}^{T} t^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^{T} u_t \\ \sum_{t=1}^{T} t u_t \end{bmatrix}$$

where the t-th observation of X, the matrix of regressors, is $[1, t]$.

(b) Use the results that $\sum_{t=1}^{T} t = T(T+1)/2$ and $\sum_{t=1}^{T} t^2 = T(T+1)(2T+1)/6$ to show that plim $(X'X/T)$ as $T \to \infty$ is not a positive definite matrix.

(c) Use the fact that

$$\begin{pmatrix} \sqrt{T}(\widehat{\alpha}_{OLS} - \alpha) \\ T\sqrt{T}(\widehat{\beta}_{OLS} - \beta) \end{pmatrix} = A(X'X)^{-1}AA^{-1}(X'u) = (A^{-1}(X'X)A^{-1})^{-1}A^{-1}(X'u)$$

$$\text{where} \qquad A = \begin{pmatrix} \sqrt{T} & 0 \\ 0 & T\sqrt{T} \end{pmatrix}$$

is the 2×2 nonsingular matrix, to show that plim $(A^{-1}(X'X)A^{-1})$ is the finite positive definite matrix

$$Q = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{pmatrix} \text{ and } \qquad A^{-1}(X'u) = \begin{pmatrix} \sum_{t=1}^{T} u_t/\sqrt{T} \\ \sum_{t=1}^{T} t u_t/T\sqrt{T} \end{pmatrix}$$

(d) Show that $z_1 = \sum_{t=1}^{T} u_t/\sqrt{T}$ is $N(0, \sigma^2)$ and $z_2 = \sum_{t=1}^{T} t u_t/T\sqrt{T}$ is $N(0, \sigma^2(T+1)(2T+1)/6T^2)$ with $\text{cov}(z_1, z_2) = (T+1)\sigma^2/2T$, so that

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \sim N\left(0, \sigma^2 \begin{pmatrix} 1 & \frac{T+1}{2T} \\ \frac{T+1}{2T} & \frac{(T+1)(2T+1)}{6T^2} \end{pmatrix}\right).$$

Conclude that as $T \to \infty$, the asymptotic distribution of $\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$ is $N(0, \sigma^2 Q)$.

(e) Using the results in parts (c) and (d), conclude that the asymptotic distribution of $\begin{pmatrix} \sqrt{T}(\widehat{\alpha}_{OLS} - \alpha) \\ T\sqrt{T}(\widehat{\beta}_{OLS} - \beta) \end{pmatrix}$ is $N(0, \sigma^2 Q^{-1})$. Since $\widehat{\beta}_{OLS}$ has the factor $T\sqrt{T}$ rather than the usual $\sqrt{T}$, it is said to be *superconsistent*. This means that not only does $(\widehat{\beta}_{OLS} - \beta)$ converge to zero in probability limits, but so does $T(\widehat{\beta}_{OLS} - \beta)$. Note that the normality assumption is not needed for this result. Using the central limit theorem, all that is needed is that u_t is White noise with finite fourth moments, see Sims, Stock and Watson (1990) or Hamilton (1994).

7. *Test of Hypothesis with a Deterministic Time Trend Model.* This is based on Hamilton (1994). In problem 6, we showed that $\widehat{\alpha}_{OLS}$ and $\widehat{\beta}_{OLS}$ converged at different rates, $\sqrt{T}$ and $T\sqrt{T}$ respectively. Despite this fact, the usual least squares t and F-statistics are asymptotically valid even when the u_t's are not Normally distributed.

   (a) Show that $s^2 = \sum_{t=1}^{T}(y_t - \widehat{\alpha}_{OLS} - \widehat{\beta}_{OLS}t)^2/(T-2)$ has plim $s^2 = \sigma^2$.

   (b) In order to test H_o; $\alpha = \alpha_o$, the usual least squares package computes

   $$t_\alpha = (\widehat{\alpha}_{OLS} - \alpha_o)/[s^2(1,0)(X'X)^{-1}(1,0)']^{1/2}$$

   where $(X'X)$ is given in problem 6. Multiply the numerator and denominator by $\sqrt{T}$ and use the results of part (c) of problem 6 to show that this t-statistic has the same asymptotic distribution as $t^*_\alpha = \sqrt{T}(\widehat{\alpha}_{OLS} - \alpha_o)/\sigma\sqrt{q^{11}}$ where q^{11} is the $(1,1)$ element of Q^{-1} defined in problem 6. t^*_α has an asymptotic $N(0,1)$ distribution using the results of part (e) in problem 6.

   (c) Similarly, to test H_o; $\beta = \beta_o$, the usual least squares package computes

   $$t_\beta = (\widehat{\beta}_{OLS} - \beta)/[s^2(0,1)(X'X)^{-1}(0,1)']^{1/2}.$$

   Multiply the numerator and denominator by $T\sqrt{T}$ and use the results of part (c) of problem 6 to show that this t-statistic has the same asymptotic distribution as $t^*_\beta = T\sqrt{T}(\widehat{\beta}_{OLS} - \beta)/\sigma\sqrt{q^{22}}$ where q^{22} is the $(2,2)$ element of Q^{-1} defined in problem 6. t^*_β has an asymptotic $N(0,1)$ distribution using the results of part (e) in problem 6.

8. *A Random Walk Model.* This is based on Fuller (1976) and Hamilton (1994). Consider the following random walk model

   $$y_t = y_{t-1} + u_t \quad t = 0, 1, \ldots, T \quad \text{where} \quad u_t \sim \text{IIN}(0, \sigma^2) \quad \text{and} \quad y_o = 0.$$

   (a) Show that y_t can be written as $y_t = u_1 + u_2 + .. + u_t$ with $E(y_t) = 0$ and $\text{var}(y_t) = t\sigma^2$ so that $y_t \sim N(0, t\sigma^2)$.

   (b) Square the random walk equation $y_t^2 = (y_{t-1} + u_t)^2$ and solve for $y_{t-1}u_t$. Sum this over $t = 1, 2, \ldots, T$ and show that

   $$\sum_{t=1}^{T} y_{t-1}u_t = (y_T^2/2) - \sum_{t=1}^{T} u_t^2/2$$

   Divide by $T\sigma^2$ and show that $\sum_{t=1}^{T} y_{t-1}u_t/T\sigma^2$ is asymptotically distributed as $(\chi^2_1 - 1)/2$. **Hint:** Use the fact that $y_T \sim N(0, T\sigma^2)$.

   (c) Using the fact that $y_{t-1} \sim N(0, (t-1)\sigma^2)$ show that $E\left(\sum_{t=1}^{T} y_{t-1}^2\right) = \sigma^2 T(T-1)/2$. **Hint:** Use the expression for $\sum_{t=1}^{T} t$ in problem 6.

   (d) Suppose we had estimated an AR(1) model rather than a random walk, i.e., $y_t = \rho y_{t-1} + u_t$ when the true $\rho = 1$. The OLS estimate is

   $$\widehat{\rho} = \sum_{t=1}^{T} y_{t-1}y_t / \sum_{t=1}^{T} y_{t-1}^2 = \rho + \sum_{t=1}^{T} y_{t-1}u_t / \sum_{t=1}^{T} y_{t-1}^2$$

   Show that

   $$\text{plim}\, T(\widehat{\rho} - \rho) = \text{plim} \frac{\sum_{t=1}^{T} y_{t-1}u_t/T\sigma^2}{\sum_{t=1}^{T} y_{t-1}^2/T^2\sigma^2} = 0$$

Note that the numerator was considered in part (b), while the denominator was considered in part (c). One can see that the asymptotic distribution of $\widehat{\rho}$ when $\rho = 1$ is a ratio of $(\chi_1^2 - 1)/2$ random variable to a non-standard distribution in the denominator which is beyond the scope of this book, see Hamilton (1994) or Fuller (1976) for further details. The object of this exercise is to show that if $\rho = 1$, $\sqrt{T}(\widehat{\rho} - \rho)$ is no longer normal as in the standard stationary least squares regression with $|\rho| < 1$. Also, to show that for the nonstationary (random walk) model, $\widehat{\rho}$ converges at a faster rate (T) than for the stationary case ($\sqrt{T}$). From part (c) it is clear that one has to divide the denominator of $\widehat{\rho}$ by T^2 rather than T to get a convergent distribution.

9. Consider the cointegration example given in (14.14) and (14.15).

   (a) Verify equations (14.16)-(14.21).

   (b) Show that the OLS estimator of β obtained by regressing C_t on Y_t is *superconsistent*, i.e., show that plim $T(\widehat{\beta}_{OLS} - \beta) \to 0$ as $T \to \infty$.

   (c) In order to check the finite sample bias of the Engle-Granger two-step estimator, let us perform the following Monte Carlo experiments: Let $\beta = 0.8$ and $\alpha = 1$ and let ρ vary over the set $\{0.6, 0.8, 0.9\}$. Also let $\sigma_{11} = \sigma_{22} = 1$, while $\sigma_{12} = 0$, and let T vary over the set $\{25, 50, 100\}$. For each of these nine experiments, generate the data on Y_t and C_t as described in (14.14) and (14.15) and estimate β and α using the Engle and Granger two-step procedure. Do 1000 replications for each experiment. Report the mean, standard deviation and MSE of α and β. Plot the bias of β versus T for various values of ρ.

References

This chapter draws on the material in Davidson and MacKinnon (1993), Maddala (1992), Hamilton (1994), Banerjee et al. (1993) and Gujarati (1995). Advanced readings include Fuller (1976) and Hamilton (1994). Easier readings include Mills (1990) and Enders (1995).

Banerjee, A., J.J. Dolado, J.W. Galbraith and D.F. Hendry (1993), *Co-Integration, Error-Correction, and The Econometric Analysis of Non-stationary Data* (Oxford University Press: Oxford).

Bierens, H.J. (2001), " Unit Roots," Chapter 29 in B.H. Baltagi (ed.) *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Bierens, H.J. and S. Guo (1993), "Testing for Stationarity and Trend Stationarity Against the Unit Root Hypothesis," *Econometric Reviews*, 12: 1-32.

Bollerslev, T. (1986), "Generalized Autoregressive Heteroskedasticity," *Journal of Econometrics*, 31: 307-327.

Box, G.E.P. and G.M. Jenkins (1970), *Time Series Analysis, Forecasting and Control* (Holden Day: San Francisco).

Box, G.E.P. and D.A. Pierce (1970), "The Distribution of Residual Autocorrelations in Auto-regressive-Integrated Moving Average Time Series Models," *Journal of American Statistical Association*, 65: 1509-1526.

Chamberlain, G. (1982), "The General Equivalence of Granger and Sims Causality," *Econometrica*, 50: 569-582.

Davidson, R. and J.G. MacKinnon (1993), *Estimation and Inference in Econometrics* (Oxford University Press: Oxford).

Dickey, D.A. and W.A. Fuller (1979), "Distribution of the Estimators for Autoregressive Time Series with A Unit Root," *Journal of the American Statistical Association*, 74: 427-431.

Dolado, J.J., J. Gonzalo and F. Marmol (2001), "Cointegration," Chapter 30 in B.H. Baltagi (ed.) *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Durlauf, S.N. and P.C.B. Phillips (1988), "Trends versus Random Walks in Time Series Analysis," *Econometrica*, 56: 1333-1354.

Enders, W. (1995), *Applied Econometric Time Series* (Wiley: New York).

Engle, R.F. (1982), "Autogressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50: 987-1007.

Engle, R.F. and C.W.J. Granger (1987), "Co-Integration and Error Correction: Representation, Estimation and Testing," *Econometrica*, 55: 251-276.

Fuller, W.A. (1976), *Introduction to Statistical Time Series* (John Wiley and Sons: New York).

Geweke, J., R. Meese and W. Dent (1983), "Comparing Alternative Tests of Causality in Temporal Systems: Analytic Results and Experimental Evidence," *Journal of Econometrics*, 21: 161-194.

Ghysels, E. and P. Perron (1993), "The Effect of Seasonal Adjustment Filters on Tests for a Unit Root," *Journal of Econometrics*, 55: 57-98.

Godfrey, L.G. (1979), "Testing the Adequacy of a Time Series Model," *Biometrika*, 66: 67-72.

Granger, C.W.J. (1969), "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," *Econometrica*, 37: 424-438.

Granger, C.W.J. (2001), "Spurious Regressions in Econometrics," Chapter 26 in B.H. Baltagi (ed.) *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Granger, C.W.J., M.L. King and H. White (1995), "Comments on Testing Economic Theories and the Use of Model Selection Criteria," *Journal of Econometrics*, 67: 173-187.

Granger, C.W.J. and P. Newbold (1974), "Spurious Regressions in Econometrics," *Journal of Econometrics*, 2: 111-120.

Gujarati, D.N. (1995), *Basic Econometrics* (McGraw Hill: New York).

Hamilton, J.D. (1994), *Time Series Analysis* (Princeton University Press: Princeton, New Jersey).

Johansen, S. (1988), "Statistical Analysis of Cointegrating Vectors," *Journal of Economic Dynamics and Control*, 12: 231-254.

Judge, G.G., R.C. Hill, W.E. Griffiths, H. Lütkepohl and T.C. Lee (1985), *The Theory and Practice of Econometrics* (John Wiley and Sons: New York).

Kwaitowski, D., P.C.B. Phillips, P. Schmidt and Y. Shin (1992), "Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root," *Journal of Econometrics*, 54: 159-178.

Leybourne, S.J. and B.P.M. McCabe (1994), "A Consistent Test for a Unit Root," *Journal of Business and Economic Statistics*, 12: 157-166.

Litterman, R.B. (1986), "Forecasting with Bayesian Vector Autoregressions-Five Years of Experience," *Journal of Business and Economic Statistics*, 4: 25-38.

Ljung, G.M. and G.E.P. Box (1978), "On a Measure of Lack of Fit in Time-Series Models," *Biometrika*, 65: 297-303.

Lütkepohl, H. (2001), "Vector Autoregressions," Chapter 32 in B.H. Baltagi (ed.) *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

MacKinnon, J.G. (1991), "Critical Values for Cointegration Tests," Ch. 13 in *Long-Run Economic Relationships: Readings in Cointegration*, eds. R.F. Engle and C.W.J. Granger (Oxford University Press: Oxford).

Maddala, G.S. (1992), *Introduction to Econometrics* (Macmillan: New York).

Mills, T.C. (1990), *Time Series Techniques for Economists* (Cambridge University Press: Cambridge).

Nelson, C.R. and C.I. Plosser (1982), "Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications," *Journal of Monetary Economics*, 10: 139-162.

Ng, S. and P. Perron (1995), "Unit Root Tests in ARMA Models With Data-Dependent Methods for the Selection of the Truncation Lag," *Journal of the American Statistical Association*, 90: 268-281.

Perron, P. (1989), "The Great Cash, The Oil Price Shock, and the Unit Root Hypothesis," *Econometrica*, 57: 1361-1401.

Phillips, P.C.B. (1986), "Understanding Spurious Regressions in Econometrics," *Journal of Econometrics*, 33: 311-340.

Phillips, P.C.B. and P. Perron (1988), "Testing for A Unit Root in Time Series Regression," *Biometrika*, 75: 335-346.

Plosser, C.I. and G.W. Shwert (1978), "Money, Income and Sunspots: Measuring Economic Relationships and the Effects of Differencing," *Journal of Monetary Economics*, 4: 637-660.

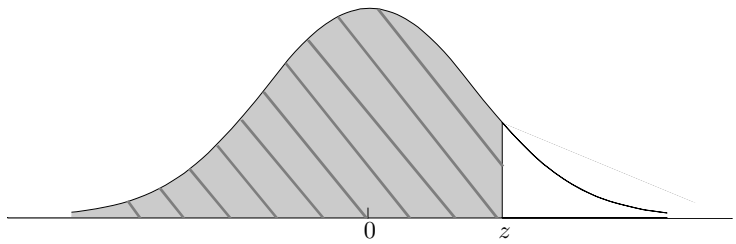
Sims, C.A. (1972), "Money, Income and Causality," *American Economic Review*, 62: 540-552.

Sims, C.A. (1980), "Macroeconomics and Reality," *Econometrica*, 48: 1-48.

Sims, C.A., J.H. Stock and M.W. Watson (1990), "Inference in Linear Time Series Models with Some Unit Roots," *Econometrica*, 58: 113-144.

Stock, J.H. and M.W. Watson (1988), "Variable Trends in Economic Time Series," *Journal of Economic Perspectives*, 2: 147-174.

Appendix

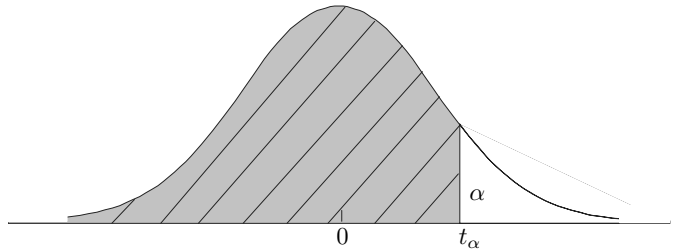


$$\Phi(1.65) = \text{pr}[z \leq 1.65] = 0.9505$$

Table A Area under the Standard Normal Distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Source: The SAS® function PROBNORM was used to generate this table.



$$\Pr[t_8 > t_\alpha = 2.306] = 0.025$$

Table B Right-Tail Critical Values for the t-Distribution

DF	**α=0.1**	**α=0.05**	**α=0.025**	**α=0.01**	**α=0.005**
1	3.0777	6.3138	12.7062	31.8205	63.6567
2	1.8856	2.9200	4.3027	6.9646	9.9248
3	1.6377	2.3534	3.1824	4.5407	5.8409
4	1.5332	2.1318	2.7764	3.7469	4.6041
5	1.4759	2.0150	2.5706	3.3649	4.0321
6	1.4398	1.9432	2.4469	3.1427	3.7074
7	1.4149	1.8946	2.3646	2.9980	3.4995
8	1.3968	1.8595	2.3060	2.8965	3.3554
9	1.3830	1.8331	2.2622	2.8214	3.2498
10	1.3722	1.8125	2.2281	2.7638	3.1693
11	1.3634	1.7959	2.2010	2.7181	3.1058
12	1.3562	1.7823	2.1788	2.6810	3.0545
13	1.3502	1.7709	2.1604	2.6503	3.0123
14	1.3450	1.7613	2.1448	2.6245	2.9768
15	1.3406	1.7531	2.1314	2.6025	2.9467
16	1.3368	1.7459	2.1199	2.5835	2.9208
17	1.3334	1.7396	2.1098	2.5669	2.8982
18	1.3304	1.7341	2.1009	2.5524	2.8784
19	1.3277	1.7291	2.0930	2.5395	2.8609
20	1.3253	1.7247	2.0860	2.5280	2.8453
21	1.3232	1.7207	2.0796	2.5176	2.8314
22	1.3212	1.7171	2.0739	2.5083	2.8188
23	1.3195	1.7139	2.0687	2.4999	2.8073
24	1.3178	1.7109	2.0639	2.4922	2.7969
25	1.3163	1.7081	2.0595	2.4851	2.7874
26	1.3150	1.7056	2.0555	2.4786	2.7787
27	1.3137	1.7033	2.0518	2.4727	2.7707
28	1.3125	1.7011	2.0484	2.4671	2.7633
29	1.3114	1.6991	2.0452	2.4620	2.7564
30	1.3104	1.6973	2.0423	2.4573	2.7500
31	1.3095	1.6955	2.0395	2.4528	2.7440
32	1.3086	1.6939	2.0369	2.4487	2.7385
33	1.3077	1.6924	2.0345	2.4448	2.7333
34	1.3070	1.6909	2.0322	2.4411	2.7284
35	1.3062	1.6896	2.0301	2.4377	2.7238
36	1.3055	1.6883	2.0281	2.4345	2.7195
37	1.3049	1.6871	2.0262	2.4314	2.7154
38	1.3042	1.6860	2.0244	2.4286	2.7116
39	1.3036	1.6849	2.0227	2.4258	2.7079
40	1.3031	1.6839	2.0211	2.4233	2.7045

Source: The SAS® function TINV was used to generate this table.

Table C Right-Tail Critical Values for the F-Distribution: Upper 5% Points

v_2/v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882	243.906	245.950	248.013	249.260	250.095	251.143
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396	19.413	19.429	19.446	19.456	19.462	19.471
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786	8.745	8.703	8.660	8.634	8.617	8.594
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.912	5.858	5.803	5.769	5.746	5.717
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.678	4.619	4.558	4.521	4.496	4.464
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.000	3.938	3.874	3.835	3.808	3.774
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.575	3.511	3.445	3.404	3.376	3.340
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347	3.284	3.218	3.150	3.108	3.079	3.043
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.073	3.006	2.936	2.893	2.864	2.826
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.913	2.845	2.774	2.730	2.700	2.661
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.788	2.719	2.646	2.601	2.570	2.531
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.687	2.617	2.544	2.498	2.466	2.426
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.604	2.533	2.459	2.412	2.380	2.339
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.534	2.463	2.388	2.341	2.308	2.266
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.475	2.403	2.328	2.280	2.247	2.204
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494	2.425	2.352	2.276	2.227	2.194	2.151
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450	2.381	2.308	2.230	2.181	2.148	2.104
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412	2.342	2.269	2.191	2.141	2.107	2.063
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378	2.308	2.234	2.155	2.106	2.071	2.026
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348	2.278	2.203	2.124	2.074	2.039	1.994
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321	2.250	2.176	2.096	2.045	2.010	1.965
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297	2.226	2.151	2.071	2.020	1.984	1.938
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275	2.204	2.128	2.048	1.996	1.961	1.914
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255	2.183	2.108	2.027	1.975	1.939	1.892
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236	2.165	2.089	2.007	1.955	1.919	1.872
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220	2.148	2.072	1.990	1.938	1.901	1.853
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204	2.132	2.056	1.974	1.921	1.884	1.836
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190	2.118	2.041	1.959	1.906	1.869	1.820
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177	2.104	2.027	1.945	1.891	1.854	1.806
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165	2.092	2.015	1.932	1.878	1.841	1.792
31	4.160	3.305	2.911	2.679	2.523	2.409	2.323	2.255	2.199	2.153	2.080	2.003	1.920	1.866	1.828	1.779
32	4.149	3.295	2.901	2.668	2.512	2.399	2.313	2.244	2.189	2.142	2.070	1.992	1.908	1.854	1.817	1.767
33	4.139	3.285	2.892	2.659	2.503	2.389	2.303	2.235	2.179	2.133	2.060	1.982	1.898	1.844	1.806	1.756
34	4.130	3.276	2.883	2.650	2.494	2.380	2.294	2.225	2.170	2.123	2.050	1.972	1.888	1.833	1.795	1.745
35	4.121	3.267	2.874	2.641	2.485	2.372	2.285	2.217	2.161	2.114	2.041	1.963	1.878	1.824	1.786	1.735
36	4.113	3.259	2.866	2.634	2.477	2.364	2.277	2.209	2.153	2.106	2.033	1.954	1.870	1.815	1.776	1.726
37	4.105	3.252	2.859	2.626	2.470	2.356	2.270	2.201	2.145	2.098	2.025	1.946	1.861	1.806	1.768	1.717
38	4.098	3.245	2.852	2.619	2.463	2.349	2.262	2.194	2.138	2.091	2.017	1.939	1.853	1.798	1.760	1.708
39	4.091	3.238	2.845	2.612	2.456	2.342	2.255	2.187	2.131	2.084	2.010	1.931	1.846	1.791	1.752	1.700
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077	2.003	1.924	1.839	1.783	1.744	1.693

Source: The SAS® function FINV was used to generate this table. v_1= numerator degrees of freedom v_2= denominator degrees of freedom

Table D Right-Tail Critical Values for the F-Distribution: Upper 1% Points

v_2/v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	25	30	40
1	4052.181	4999.500	5403.352	5624.583	5763.650	5858.986	5928.356	5981.070	6022.473	6055.847	6106.321	6157.285	6208.730	6239.825	6260.649	6286.782
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399	99.416	99.433	99.449	99.459	99.466	99.474
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229	27.052	26.872	26.690	26.579	26.505	26.411
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.374	14.198	14.020	13.911	13.838	13.745
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.888	9.722	9.553	9.449	9.379	9.291
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874	7.718	7.559	7.396	7.296	7.229	7.143
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620	6.469	6.314	6.155	6.058	5.992	5.908
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.667	5.515	5.359	5.263	5.198	5.116
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.111	4.962	4.808	4.713	4.649	4.567
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849	4.706	4.558	4.405	4.311	4.247	4.165
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.397	4.251	4.099	4.005	3.941	3.860
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296	4.155	4.010	3.858	3.765	3.701	3.619
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100	3.960	3.815	3.665	3.571	3.507	3.425
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939	3.800	3.656	3.505	3.412	3.348	3.266
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.666	3.522	3.372	3.278	3.214	3.132
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691	3.553	3.409	3.259	3.165	3.101	3.018
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593	3.455	3.312	3.162	3.068	3.003	2.920
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508	3.371	3.227	3.077	2.983	2.919	2.835
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434	3.297	3.153	3.003	2.909	2.844	2.761
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.231	3.088	2.938	2.843	2.778	2.695
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310	3.173	3.030	2.880	2.785	2.720	2.636
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258	3.121	2.978	2.827	2.733	2.667	2.583
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211	3.074	2.931	2.781	2.686	2.620	2.535
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168	3.032	2.889	2.738	2.643	2.577	2.492
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129	2.993	2.850	2.699	2.604	2.538	2.453
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094	2.958	2.815	2.664	2.569	2.503	2.417
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062	2.926	2.783	2.632	2.536	2.470	2.384
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032	2.896	2.753	2.602	2.506	2.440	2.354
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005	2.868	2.726	2.574	2.478	2.412	2.325
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979	2.843	2.700	2.549	2.453	2.386	2.299
31	7.530	5.362	4.484	3.993	3.675	3.449	3.281	3.149	3.043	2.955	2.820	2.677	2.525	2.429	2.362	2.275
32	7.499	5.336	4.459	3.969	3.652	3.427	3.258	3.127	3.021	2.934	2.798	2.655	2.503	2.406	2.340	2.252
33	7.471	5.312	4.437	3.948	3.630	3.406	3.238	3.106	3.000	2.913	2.777	2.634	2.482	2.386	2.319	2.231
34	7.444	5.289	4.416	3.927	3.611	3.386	3.218	3.087	2.981	2.894	2.758	2.615	2.463	2.366	2.299	2.211
35	7.419	5.268	4.396	3.908	3.592	3.368	3.200	3.069	2.963	2.876	2.740	2.597	2.445	2.348	2.281	2.193
36	7.396	5.248	4.377	3.890	3.574	3.351	3.183	3.052	2.946	2.859	2.723	2.580	2.428	2.331	2.263	2.175
37	7.373	5.229	4.360	3.873	3.558	3.334	3.167	3.036	2.930	2.843	2.707	2.564	2.412	2.315	2.247	2.159
38	7.353	5.211	4.343	3.858	3.542	3.319	3.152	3.021	2.915	2.828	2.692	2.549	2.397	2.299	2.232	2.143
39	7.333	5.194	4.327	3.843	3.528	3.305	3.137	3.006	2.901	2.814	2.678	2.535	2.382	2.285	2.217	2.128
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801	2.665	2.522	2.369	2.271	2.203	2.114

Source: The SAS® function FINV was used to generate this table. v_1= numerator degrees of freedom v_2 = denominator degrees of freedom

Table E Right-Tail Critical Values for the Chi-Square Distribution $\Pr[\chi^2_5 > 11.0705] = 0.05$

v	.995	.990	.975	.950	.90	.50	.10	.05	.025	.01	.005
1	0.00004	0.00016	0.00098	0.00393	0.01579	0.45494	2.70554	3.84146	5.02389	6.63490	7.87944
2	0.01003	0.02010	0.05064	0.10259	0.21072	1.38629	4.60517	5.99146	7.37776	9.21034	10.5966
3	0.07172	0.11483	0.21580	0.35185	0.58437	2.36597	6.25139	7.81473	9.34840	11.3449	12.8382
4	0.20699	0.29711	0.48442	0.71072	1.06362	3.35669	7.77944	9.48773	11.1433	13.2767	14.8603
5	0.41174	0.55430	0.83121	1.14548	1.61031	4.35146	9.23636	11.0705	12.8325	15.0863	16.7496
6	0.67573	0.87209	1.23734	1.63538	2.20413	5.34812	10.6446	12.5916	14.4494	16.8119	18.5476
7	0.98926	1.23904	1.68987	2.16735	2.83311	6.34581	12.0170	14.0671	16.0128	18.4753	20.2777
8	1.34441	1.64650	2.17973	2.73264	3.48954	7.34412	13.3616	15.5073	17.5345	20.0902	21.9550
9	1.73493	2.08790	2.70039	3.32511	4.16816	8.34283	14.6837	16.9190	19.0228	21.6660	23.5894
10	2.15586	2.55821	3.24697	3.94030	4.86518	9.34182	15.9872	18.3070	20.4832	23.2093	25.1882
11	2.60322	3.05348	3.81575	4.57481	5.57778	10.3410	17.2750	19.6751	21.9200	24.7250	26.7568
12	3.07382	3.57057	4.40379	5.22603	6.30380	11.3403	18.5493	21.0261	23.3367	26.2170	28.2995
13	3.56503	4.10692	5.00875	5.89186	7.04150	12.3398	19.8119	22.3620	24.7356	27.6882	29.8195
14	4.07467	4.66043	5.62873	6.57063	7.78953	13.3393	21.0641	23.6848	26.1189	29.1412	31.3193
15	4.60092	5.22935	6.26214	7.26094	8.54676	14.3389	22.3071	24.9958	27.4884	30.5779	32.8013
16	5.14221	5.81221	6.90766	7.96165	9.31224	15.3385	23.5418	26.2962	28.8454	31.9999	34.2672
17	5.69722	6.40776	7.56419	8.67176	10.0852	16.3382	24.7690	27.5871	30.1910	33.4087	35.7185
18	6.26480	7.01491	8.23075	9.39046	10.8649	17.3379	25.9894	28.8693	31.5264	34.8053	37.1565
19	6.84397	7.63273	8.90652	10.1170	11.6509	18.3377	27.2036	30.1435	32.8523	36.1909	38.5823
20	7.43384	8.26040	9.59078	10.8508	12.4426	19.3374	28.4120	31.4104	34.1696	37.5662	39.9968
21	8.03365	8.89720	10.2829	11.5913	13.2396	20.3372	29.6151	32.6706	35.4789	38.9322	41.4011
22	8.64272	9.54249	10.9823	12.3380	14.0415	21.3370	30.8133	33.9244	36.7807	40.2894	42.7957
23	9.26042	10.1957	11.6886	13.0905	14.8480	22.3369	32.0069	35.1725	38.0756	41.6384	44.1813
24	9.88623	10.8564	12.4012	13.8484	15.6587	23.3367	33.1962	36.4150	39.3641	42.9798	45.5585
25	10.5197	11.5240	13.1197	14.6114	16.4734	24.3366	34.3816	37.6525	40.6465	44.3141	46.9279
26	11.1602	12.1981	13.8439	15.3792	17.2919	25.3365	35.5632	38.8851	41.9232	45.6417	48.2899
27	11.8076	12.8785	14.5734	16.1514	18.1139	26.3363	36.7412	40.1133	43.1945	46.9629	49.6449
28	12.4613	13.5647	15.3079	16.9279	18.9392	27.3362	37.9159	41.3371	44.4608	48.2782	50.9934
29	13.1211	14.2565	16.0471	17.7084	19.7677	28.3361	39.0875	42.5570	45.7223	49.5879	52.3356
30	13.7867	14.9535	16.7908	18.4927	20.5992	29.3360	40.2560	43.7730	46.9792	50.8922	53.6720
31	14.4578	15.6555	17.5387	19.2806	21.4336	30.3359	41.4217	44.9853	48.2319	52.1914	55.0027
32	15.1340	16.3622	18.2908	20.0719	22.2706	31.3359	42.5847	46.1943	49.4804	53.4858	56.3281
33	15.8153	17.0735	19.0467	20.8665	23.1102	32.3358	43.7452	47.3999	50.7251	54.7755	57.6484
34	16.5013	17.7891	19.8063	21.6643	23.9523	33.3357	44.9032	48.6024	51.9660	56.0609	58.9639
35	17.1918	18.5089	20.5694	22.4650	24.7967	34.3356	46.0588	49.8018	53.2033	57.3421	60.2748
36	17.8867	19.2327	21.3359	23.2686	25.6433	35.3356	47.2122	50.9985	54.4373	58.6192	61.5812
37	18.5858	19.9602	22.1056	24.0749	26.4921	36.3355	48.3634	52.1923	55.6680	59.8925	62.8833
38	19.2889	20.6914	22.8785	24.8839	27.3430	37.3355	49.5126	53.3835	56.8955	61.1621	64.1814
39	19.9959	21.4262	23.6543	25.6954	28.1958	38.3354	50.6598	54.5722	58.1201	62.4281	65.4756
40	20.7065	22.1643	24.4330	26.5093	29.0505	39.3353	51.8051	55.7585	59.3417	63.6907	66.7660

Source: The SAS® function CINV was used to generate this table. v denotes the degrees of freedom.

List of Figures

List of Tables

Index