ECONOMETRICS

Bruce E. Hansen

University of Wisconsin
www.ssc.wisc.edu/~bhansen

Revised: January 2006
Comments Welcome

Contents

Chapter 1

Introduction

Econometrics is the study of estimation and inference for economic models using economic data. Econometric theory concerns the study and development of tools and methods for applied econometric applications. Applied econometrics concerns the application of these tools to economic data.

1.1 Economic Data

An econometric study requires data for analysis. The quality of the study will be largely determined by the data available. There are three major types of economic data sets: cross-sectional, time-series, and panel. They are distinguished by the dependence structure across observations.

Cross-sectional data sets are characterized by mutually independent observations. Surveys are a typical source for cross-sectional data. The individuals surveyed may be persons, households, or corporations.

Time-series data is indexed by time. Typical examples include macroeconomic aggregates, prices and interest rates. This type of data is characterized by serial dependence.

Panel data combines elements of cross-section and time-series. These data sets consist surveys of a set of individuals, repeated over time. Each individual (person, household or corporation) is surveyed on multiple occasions.

1.2 Observational Data

A common econometric question is to quantify the impact of one set of variables on another variable. For example, a concern in labor economics is the returns to schooling – the change in earnings induced by increasing a worker's education, holding other variables constant. Another issue of interest is the earnings gap between men and women.

Ideally, we would use **experimental** data to answer these questions. To measure the returns to schooling, an experiment might randomly divide children into groups, mandate different levels of education to the different groups, and then follow the children's wage path as they mature and enter the labor force. The differences between the groups could be attributed to the different levels of education. However, experiments such as this are infeasible, even immoral!

Instead, most economic data is **observational**. To continue the above example, what we observe (through data collection) is the level of a person's education and their wage. We can measure the joint distribution of these variables, and assess the joint dependence. But we cannot infer causality, as we are not able to manipulate one variable to see the direct effect on the other.

For example, a person's level of education is (at least partially) determined by that person's choices and their achievement in education. These factors are likely to be affected by their personal abilities and attitudes towards work. The fact that a person is highly educated suggests a high level of ability. This is an alternative explanation for an observed positive correlation between educational levels and wages. High ability individuals do better in school, and therefore choose to attain higher levels of education, and their high ability is the fundamental reason for their high wages. The point is that multiple explanations are consistent with a positive correlation between schooling levels and education. Knowledge of the joint distibution cannot distinguish between these explanations.

This discussion means that causality cannot be infered from observational data alone. Causal inference requires identification, and this is based on strong assumptions. We will return to a discussion of some of these issues in Chapter 11.

1.3 Random Sample

Typically, an econometrician has data

$$\{(y_1, x_1), (y_2, x_2), ..., (y_i, x_i), ..., (y_n, x_n)\} = \{(y_i, x_i) : i = 1, ..., n\}$$

where each pair $\{y_i, x_i\} \in R \times R^k$ is an **observation** on an individual (e.g., household or firm). We call these observations the **sample.**

If the data is **cross-sectional** (each observation is a different individual) it is often reasonable to assume they are mutually independent. If the data is randomly gathered, it is reasonable to model each observation as a random draw from the same probability distribution. In this case the data are **independent and identically distributed**, or **iid**. We call this a **random sample**. Sometimes the independent part of the label iid is misconstrued. It is not a statement about the relationship between y_i and x_i. Rather it means that the pair (y_i, x_i) is independent of the pair (y_j, x_j) for $i \neq j$.

The random variables (y_i, x_i) have a distribution F which we call the **population**. This "population" is infinitely large. This abstraction can a source of confusion as it does not correspond to a physical population in the real world. The distribution F is unknown, and the goal of statistical inference is to learn about features of F from the sample.

At this point in our analysis it is unimportant whether the observations y_i and x_i come from continuous or discrete distributions. For example, many regressors in econometric practice are binary, taking on only the values 0 and 1, and are typically called **dummy variables**.

1.4 Economic Data

Fortunately for economists, the development of the internet has provided a convenient forum for dissemination of economic data. Many large-scale economic datasets are available without charge from governmental agencies. An excellent starting point is the Resources for Economists Data Links, available at `http://rfe.wustl.edu/Data/index.html`.

Some other excellent data sources are listed below.
Bureau of Labor Statistics: `http://www.bls.gov/`
Federal Reserve Bank of St. Louis: `http://research.stlouisfed.org/fred2/`
Board of Governors of the Federal Reserve System: `http://www.federalreserve.gov/releases/`
National Bureau of Economic Research: `http://www.nber.org/`
US Census: `http://www.census.gov/econ/www/`

Current Population Survey (CPS): http://www.bls.census.gov/cps/cpsmain.htm
Survey of Income and Program Participation (SIPP): http://www.sipp.census.gov/sipp/
Panel Study of Income Dynamics (PSID): http://psidonline.isr.umich.edu/
U.S. Bureau of Economic Analysis: http://www.bea.doc.gov/
CompuStat: http://www.compustat.com/www/
International Financial Statistics (IFS): http://ifs.apdi.net/imf/

Chapter 2

Matrix Algebra

This chapter reviews the essential components of matrix algebra.

2.1 Terminology

A **scalar** a is a single number.

A **vector** a is a $k \times 1$ list of numbers, typically arranged in a column. We write this as

$$a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix}$$

Equivalently, a vector a is an element of Euclidean k space, hence $a \in R^k$. If $k = 1$ then a is a scalar.

A **matrix** A is a $k \times r$ rectangular array of numbers, written as

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kr} \end{bmatrix} = [a_{ij}]$$

By convention a_{ij} refers to the $i'th$ row and $j'th$ column of A. If $r = 1$ or $k = 1$ then A is a vector. If $r = k = 1$, then A is a scalar.

A matrix can be written as a set of column vectors or as a set of row vectors. That is,

$$A = \begin{bmatrix} a_1 & a_2 & \cdots & a_r \end{bmatrix} = \begin{bmatrix} \alpha_1' \\ \alpha_2' \\ \vdots \\ \alpha_k' \end{bmatrix}$$

where

$$a_i = \begin{bmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{ki} \end{bmatrix}$$

are column vectors and

$$\alpha_j' = \begin{bmatrix} a_{j1} & a_{j2} & \cdots & a_{jr} \end{bmatrix}$$

are row vectors.

The **transpose** of a matrix, denoted $B = A'$, is obtained by flipping the matrix on its diagonal.

$$B = A' = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{k1} \\ a_{12} & a_{22} & \cdots & a_{k2} \\ \vdots & \vdots & & \vdots \\ a_{1r} & a_{2r} & \cdots & a_{kr} \end{bmatrix}$$

Thus $b_{ij} = a_{ji}$ for all i and j. Note that if A is $k \times r$, then A' is $r \times k$. If a is a $k \times 1$ vector, then a' is a $1 \times k$ row vector.

A matrix is **square** if $k = r$. A square matrix is **symmetric** if $A = A'$, which implies $a_{ij} = a_{ji}$. A square matrix is **diagonal** if the only non-zero elements appear on the diagonal, so that $a_{ij} = 0$ if $i \neq j$. A square matrix is **upper (lower) diagonal** if all elements below (above) the diagonal equal zero.

A **partitioned matrix** takes the form

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1r} \\ A_{21} & A_{22} & \cdots & A_{2r} \\ \vdots & \vdots & & \vdots \\ A_{k1} & A_{k2} & \cdots & A_{kr} \end{bmatrix}$$

where the A_{ij} denote matrices, vectors and/or scalars.

2.2 Matrix Multiplication

If a and b are both $k \times 1$, then their inner product is

$$a'b = a_1b_1 + a_2b_2 + \cdots + a_kb_k = \sum_{j=1}^{k} a_jb_j$$

Note that $a'b = b'a$.

If A is $k \times r$ and B is $r \times s$, then we define their product AB by writing A as a set of row vectors and B as a set of column vectors (each of length r). Then

$$\begin{aligned} AB &= \begin{bmatrix} a_1' \\ a_2' \\ \vdots \\ a_k' \end{bmatrix} \begin{bmatrix} b_1 & b_2 & \cdots & b_s \end{bmatrix} \\ &= \begin{bmatrix} a_1'b_1 & a_1'b_2 & \cdots & a_1'b_s \\ a_2'b_1 & a_2'b_2 & \cdots & a_2'b_s \\ \vdots & \vdots & & \vdots \\ a_k'b_1 & a_k'b_2 & \cdots & a_k'b_s \end{bmatrix} \end{aligned}$$

When the number of columns of A equals the number of rows of B, we say that A and B, or the product AB, are **conformable**, and this is the only case where this product is defined.

An alternative way to write the matrix product is to use matrix partitions. For example,

$$\begin{aligned} AB &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \\ &= \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} AB &= \begin{bmatrix} A_1 & A_2 & \cdots & A_r \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_r \end{bmatrix} \\ &= A_1B_1 + A_2B_2 + \cdots + A_rB_r \\ &= \sum_{j=1}^{r} A_rB_r \end{aligned}$$

An important diagonal matrix is the **identity matrix**, which has ones on the diagonal. A $k \times k$ identity matrix is denoted as

$$I_k = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Important properties are that if A is $k \times r$, then $AI_r = A$ and $I_kA = A$.

We say that two vectors a and b are **orthogonal** if $a'b = 0$. The columns of a $k \times r$ matrix A, $r \leq k$, are said to be orthogonal if $A'A = I_r$. A square matrix A is called orthogonal if $A'A = I_k$.

2.3 Trace

The **trace** of a $k \times k$ square matrix A is the sum of its diagonal elements

$$\operatorname{tr}(A) = \sum_{i=1}^{k} a_{ii}$$

Some straightforward properties for square matrices A and B are

$$\begin{aligned} \operatorname{tr}(cA) &= c \operatorname{tr}(A) \\ \operatorname{tr}\left(A'\right) &= \operatorname{tr}(A) \\ \operatorname{tr}(A + B) &= \operatorname{tr}(A) + \operatorname{tr}(B) \\ \operatorname{tr}(I_k) &= k. \end{aligned}$$

Also, for $k \times r$ A and $r \times k$ B we have

$$\operatorname{tr}(AB) = \operatorname{tr}(BA)$$

The can be seen since

$$\begin{aligned}
\operatorname{tr}(AB) &= \operatorname{tr}\begin{bmatrix} a_1'b_1 & a_1'b_2 & \cdots & a_1'b_k \\ a_2'b_1 & a_2'b_2 & \cdots & a_2'b_k \\ \vdots & \vdots & & \vdots \\ a_k'b_1 & a_k'b_2 & \cdots & a_k'b_k \end{bmatrix} \\
&= \sum_{i=1}^{k} a_i'b_i \\
&= \sum_{i=1}^{k} b_i'a_i \\
&= \operatorname{tr}(BA).
\end{aligned}$$

2.4 Inverse

A $k \times k$ matrix A has **full rank,** or is **nonsingular**, if there is no $c \neq 0$ such that $Ac = 0$. In this case there exists a unique matrix B such that $AB = BA = I_k$. This matrix is called the **inverse** of A and is denoted by A^{-1}. For non-singular A and C, some properties include

$$\begin{aligned}
AA^{-1} &= A^{-1}A = I_k \\
\left(A^{-1}\right)' &= \left(A'\right)^{-1} \\
(AC)^{-1} &= C^{-1}A^{-1} \\
(A+C)^{-1} &= A^{-1}\left(A^{-1}+C^{-1}\right)^{-1}C^{-1} \\
A^{-1}-(A+C)^{-1} &= A^{-1}\left(A^{-1}+C^{-1}\right)A^{-1} \\
(A+BCD)^{-1} &= A^{-1}-A^{-1}BC\left(C+CDA^{-1}BC\right)CDA^{-1}
\end{aligned} \tag{2.1}$$

Also, if A is an orthogonal matrix, then $A^{-1} = A$.

The following fact about inverting partitioned matrices is sometimes useful. If $A - BD^{-1}C$ and $D - CA^{-1}B$ are non-singular, then

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} \left(A-BD^{-1}C\right)^{-1} & -\left(A-BD^{-1}C\right)^{-1}BD^{-1} \\ -\left(D-CA^{-1}B\right)^{-1}CA^{-1} & \left(D-CA^{-1}B\right)^{-1} \end{bmatrix}. \tag{2.2}$$

Even if a matrix A does not possess an inverse, we can still define a **generalized inverse** A^- as a matrix which satisfies

$$AA^-A = A. \tag{2.3}$$

The matrix A^- is not necessarily unique. The **Moore-Penrose generalized inverse** A^- satisfies (2.3) plus the following three conditions

$$\begin{aligned}
&A^-AA^- = A^- \\
&AA^- \text{ is symmetric} \\
&A^-A \text{ is symmetric}
\end{aligned}$$

For any matrix A, the Moore-Penrose generalized inverse A^- exists and is unique.

2.5 Eigenvalues

The characteristic equation of a square matrix A is

$$\det(A - \lambda I_k) = 0.$$

The left side is a polynomial of degree k in λ so it has exactly k roots, which are not necessarily distinct and may be real or complex. They are called the **latent roots** or **characteristic roots** or **eigenvalues** of A. If λ_i is an eigenvalue of A, then $A - \lambda_i I_k$ is singular so there exists a non-zero vector h_i such that

$$(A - \lambda_i I_k) h_i = 0$$

The vector h_i is called a **latent vector** or **characteristic vector** or **eigenvector** of A corresponding to λ_i.

We now state some useful properties. Let λ_i and h_i, $i = 1, ..., k$ denote the k eigenvalues and eigenvectors of a square matrix A. Let Λ be a diagonal matrix with the characteristic roots in the diagonal, and let $H = [h_1 \cdots h_k]$.

- $\det(A) = \prod_{i=1}^{k} \lambda_i$
- $\mathrm{tr}(A) = \sum_{i=1}^{k} \lambda_i$
- A is non-singular if and only if all its characteristic roots are non-zero.
- If A has distinct characteristic roots, there exists a nonsingular matrix P such that $A = P^{-1}\Lambda P$ and $PAP^{-1} = \Lambda$.
- If A is symmetric, then $A = H\Lambda H'$ and $H'AH = \Lambda$, and the characteristic roots are all real.
- The characteristic roots of A^{-1} are λ_1^{-1}, λ_2^{-1}, ..., λ_k^{-1}.

The decomposition $A = H\Lambda H'$ is called the **spectral decomposition** of a matrix.

2.6 Rank and Positive Definiteness

The **rank** of a square matrix is the number of its non-zero characteristic roots.

We say that a square matrix A is **positive semi-definite** if for all non-zero c, $c'Ac \geq 0$. This is written as $A \geq 0$. We say that A is **positive definite** if for all non-zero c, $c'Ac > 0$. This is written as $A > 0$.

If $A = G'G$, then A is positive semi-definite. (For any $c \neq 0$, $c'Ac = \alpha'a \geq 0$ where $\alpha = Gc$.)

If A is positive definite, then A is non-singular and A^{-1} exists. Furthermore, $A^{-1} > 0$.

We say that X is $n \times k$, $k < n$, has **full rank** k if there is no non-zero c such that $Xc = 0$. In this case, $X'X$ is symmetric and positive definite.

If A is symmetric, then $A > 0$ if and only if all its characteristic roots are positive.

If $A > 0$ we can find a matrix B such that $A = BB'$. We call B a **matrix square root** of A. The matrix B need not be unique. One way to construct B is to use the spectral decomposition $A = H\Lambda H'$ where Λ is diagonal, and then set $B = H\Lambda^{1/2}$.

A square matrix A is **idempotent** if $AA = A$. If A is also symmetric (most idempotent matrices are) then all its characteristic roots equal either zero or one. To see this, note that we can write $A = H\Lambda H'$ where H is orthogonal and Λ contains the (real) characteristic roots. Then

$$A = AA = H\Lambda H' H\Lambda H' = H\Lambda^2 H'.$$

By the uniqueness of the characteristic roots, we deduce that $\Lambda^2 = \Lambda$ and $\lambda_i^2 = \lambda_i$ for $i = 1, ..., k$. Hence they must equal either 0 or 1. It follows that the spectral decomposition of A takes the form

$$M = H \begin{bmatrix} I_{n-k} & 0 \\ 0 & 0 \end{bmatrix} H' \tag{2.4}$$

with $H'H = I_n$. Additionally, $\text{tr}(A) = rank(A)$.

2.7 Matrix Calculus

Let $x = (x_1, ..., x_k)$ be $k \times 1$ and $g(x) = g(x_1, ..., x_k) : R^k \to R$. The vector derivative is

$$\frac{\partial}{\partial x} g(x) = \begin{pmatrix} \frac{\partial}{\partial x_1} g(x) \\ \vdots \\ \frac{\partial}{\partial x_k} g(x) \end{pmatrix}$$

and

$$\frac{\partial}{\partial x'} g(x) = \begin{pmatrix} \frac{\partial}{\partial x_1} g(x) & \cdots & \frac{\partial}{\partial x_k} g(x) \end{pmatrix}.$$

Some properties are now summarized.

- $\frac{\partial}{\partial x}(a'x) = \frac{\partial}{\partial x}(x'a) = a$
- $\frac{\partial}{\partial x'}(Ax) = A$
- $\frac{\partial}{\partial x}(x'Ax) = (A + A')x$
- $\frac{\partial^2}{\partial x \partial x'}(x'Ax) = A + A'$

2.8 Determinant

The **determinant** is defined for square matrices.

If A is 2×2, then its determinant is $\det A = a_{11}a_{22} - a_{12}a_{21}$.

For a general $k \times k$ matrix $A = [a_{ij}]$, we can define the determinant as follows. Let $\pi = (j_1, ..., j_k)$ denote a permutation of $(1, ..., k)$. There are $k!$ such permutations. There is a unique count of the number of inversions of the indices of such permutations (relative to the natural order $(1, ..., k)$), and let $\varepsilon_\pi = +1$ if this count is even and $\varepsilon_\pi = -1$ if the count is odd. Then

$$\det A = \sum_\pi \varepsilon_\pi a_{1j_1} a_{2j_2} \cdots a_{kj_k}$$

Some properties include

- $\det A = \det A'$
- $\det(\alpha A) = \alpha^k \det A$
- $\det(AB) = (\det A)(\det B)$
- $\det\left(A^{-1}\right) = (\det A)^{-1}$

- $\det \begin{bmatrix} A & B \\ C & D \end{bmatrix} = (\det D) \det \left(A - BD^{-1}C\right)$ if $\det D \neq 0$
- $\det A \neq 0$ if and only if A is nonsingular.
- If A is triangular (upper or lower), then $\det A = \prod_{i=1}^{k} a_{ii}$
- If A is orthogonal, then $\det A = \pm 1$

2.9 Kronecker Products and the Vec Operator

Let $A = [a_1 \ a_2 \ \cdots \ a_n] = [a_{ij}]$ be $m \times n$. The **vec** of A, denoted by $\operatorname{vec}(A)$, is the $mn \times 1$ vector

$$\operatorname{vec}(A) = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}.$$

Let B be any matrix. The **Kronecker product** of A and B, denoted $A \otimes B$, is the matrix

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{bmatrix}.$$

Some important properties are now summarized. These results hold for matrices for which all matrix multiplications are conformable.

- $(A + B) \otimes C = A \otimes C + B \otimes C$
- $(A \otimes B)(C \otimes D) = AC \otimes BD$
- $A \otimes (B \otimes C) = (A \otimes B) \otimes C$
- $(A \otimes B)' = A' \otimes B'$
- $\operatorname{tr}(A \otimes B) = \operatorname{tr}(A) \operatorname{tr}(B)$
- If A is $m \times m$ and B is $n \times n$, $\det(A \otimes B) = (\det(A))^n (\det(B))^m$
- $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$
- If $A > 0$ and $B > 0$ then $A \otimes B > 0$
- $\operatorname{vec}(ABC) = (C' \otimes A) \operatorname{vec}(B)$
- $\operatorname{tr}(ABCD) = \operatorname{vec}(D')' (C' \otimes A) \operatorname{vec}(B)$

Chapter 3

Regression and Projection

The most commonly applied econometric tool is regression. This is used when the goal is to quantify the impact of one set of variables on another variable. In this context we partition the observations into the pair (y_i, x_i) where y_i is a scalar (real-valued) and x_i is a vector. We call y_i the **dependent variable.** We call x_i alternatively the **regressor**, the **conditioning variable**, or the **covariates**. We list the elements of x_i in the vector

$$x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ki} \end{pmatrix}. \tag{3.1}$$

3.1 Conditional Mean

To study how the distribution of y_i varies with the variables x_i in the population, we can focus on $f(y \mid x)$, the conditional density of y_i given x_i.

To illustrate, Figure 3.1 displays the density[1] of hourly wages for men and women, from the population of white non-military wage earners with a college degree and 10-15 years of potential work experience. These are conditional density functions – the density of hourly wages conditional on race, gender, education and experience. The two density curves show the effect of gender on the distribution of wages, holding the other variables constant.

While it is easy to observe that the two densities are unequal, it is useful to have numerical measures of the difference. An important summary measure is the **conditional mean**

$$m(x) = E(y_i \mid x_i = x) = \int_{-\infty}^{\infty} y f(y \mid x)\, dy. \tag{3.2}$$

In general, $m(x)$ can take any form, and exists so long as $E|y_i| < \infty$. In the example presented in Figure 3.1, the mean wage for men is \$27.22, and that for women is \$20.73. These are indicated in Figure 3.1 by the arrows drawn to the x-axis.

Take a closer look at the density functions displayed in Figure 3.1. You can see that the right tail of then density is much thicker than the left tail. These are asymmetric (skewed) densities, which is a common feature of wage distributions. When a distribution is skewed, the mean is not necessarily a good summary of the central tendency. In this context it is often convenient

[1] These are nonparametric density estimates using a Gaussian kernel with the bandwidth selected by cross-validation. See Chapter 16. The data are from the 2004 Current Population Survey

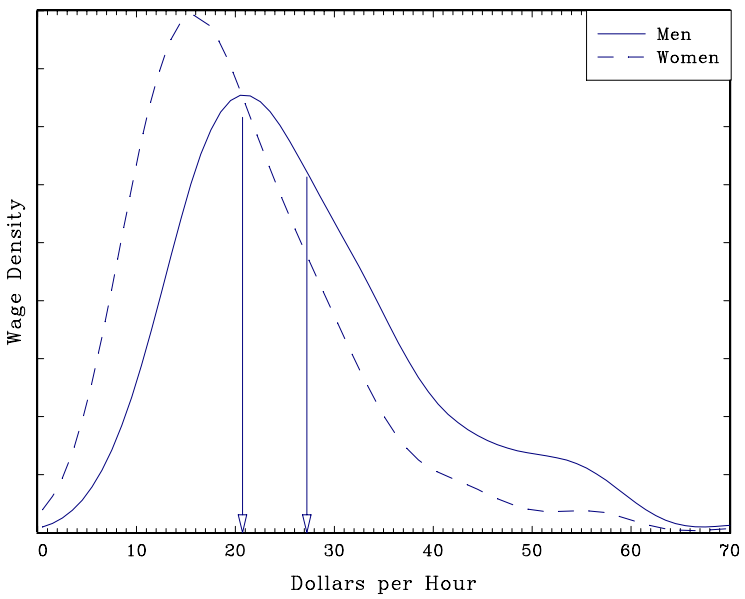

Figure 3.1: Wage Densities for White College Grads with 10-15 Years Work Experience

to transform the data by taking the (natural) logarithm. Figure 3.2 shows the density of log hourly wages for the same population, with mean log hourly wages drawn in with the arrows. The difference in the log mean wage between men and women is 0.30, which implies a 30% average wage difference for this population. This is a more robust measure of the typical wage gap between men and women than the difference in the untransformed wage means. For this reason, wage regressions typically use log wages as a dependent variable rather than the level of wages.

The comparison in Figure 3.1 is facilitated by the fact that the control variable (gender) is discrete. When the distribution of the control variable is continuous, then comparisons become more complicated. To illustrate, Figure 3.3 displays a scatter plot[2] of log wages against education levels. Assuming for simplicity that this is the true joint distribution, the solid line displays the conditional expectation of log wages varying with education. The conditional expectation function is close to linear; the dashed line is a linear projection approximation which will be discussed in the Section 3.5. The main point to be learned from Figure 3.3 is how the conditional expectation describes an important feature of the conditional distribution. Of particular interest to graduate students may be the observation that difference between a B.A. and a Ph.D. degree in mean log hourly wages is 0.36, implying an average 36% difference in wage levels.

3.2 Regression Equation

The regression error e_i is defined to be the difference between y_i and its conditional mean (3.2) evaluated at the observed value of x_i:

$$e_i = y_i - m(x_i).$$

By construction, this yields the formula

$$y_i = m(x_i) + e_i. \tag{3.3}$$

[2] White non-military male wage earners with 10-15 years of potential work experience.

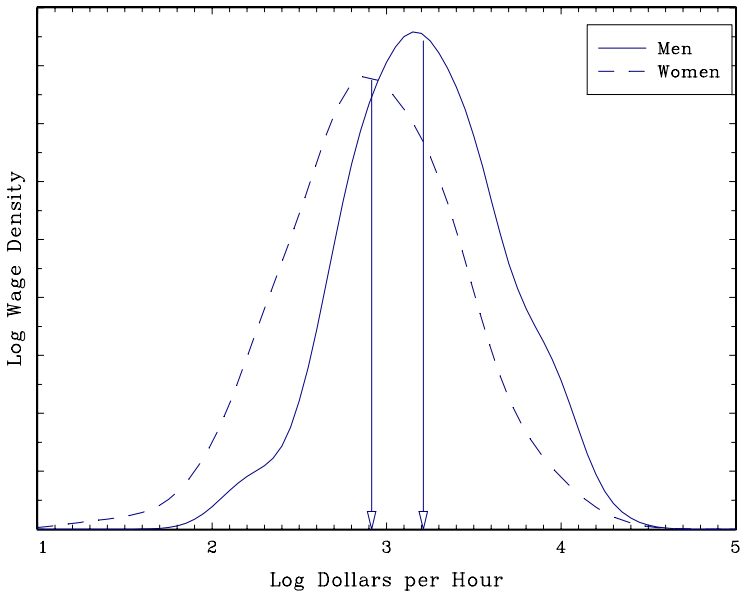

Figure 3.2: Log Wage Densities

Theorem 3.2.1 *Properties of the regression error* e_i

1. $E\left(e_i \mid x_i\right) = 0.$
2. $E(e_i) = 0.$
3. $E\left(h(x_i)e_i\right) = 0$ *for any function* $h\left(\cdot\right).$
4. $E(x_i e_i) = 0.$

To show the first statement, by the definition of e_i and the linearity of conditional expectations,

$$\begin{aligned} E\left(e_i \mid x_i\right) &= E\left(\left(y_i - m(x_i)\right) \mid x_i\right) \\ &= E\left(y_i \mid x_i\right) - E\left(m(x_i) \mid x_i\right) \\ &= m(x_i) - m(x_i) \\ &= 0. \end{aligned}$$

The remaining parts of the Theorem are left as an exercise.

The equations

$$\begin{aligned} y_i &= m(x_i) + e_i \\ E\left(e_i \mid x_i\right) &= 0. \end{aligned}$$

are often stated jointly as the regression framework. It is important to understand that this is a framework, not a model, because no restrictions have been placed on the joint distribution of the data. These equations hold true by definition. A regression model imposes further restrictions on the joint distribution; most typically, restrictions on the permissible class of regression functions $m(x)$.

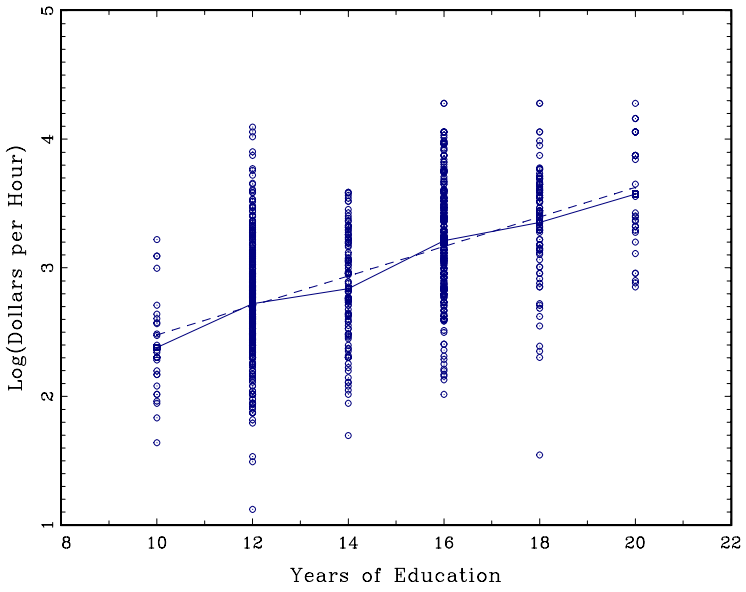

Figure 3.3: Conditional Mean of Wages Given Education

The conditional mean also has the property of being the the best predictor of y_i, in the sense of achieving the lowest mean squared error. To see this, let $g(x)$ be an arbitrary predictor of y_i given $x_i = x$. The expected squared error using this prediction function is

$$\begin{aligned} E\left(y_i - g(x_i)\right)^2 &= E\left(e_i + m(x_i) - g(x_i)\right)^2 \\ &= Ee_i^2 + 2E\left(e_i\left(m(x_i) - g(x_i)\right)\right) + E\left(m(x_i) - g(x_i)\right)^2 \\ &= Ee_i^2 + E\left(m(x_i) - g(x_i)\right)^2 \\ &\geq Ee_i^2 \end{aligned}$$

where the second equality uses Theorem 3.2.1.3. The right-hand-side is minimized by setting $g(x) = m(x)$. Thus the mean squared error is minimized by the conditional mean.

3.3 Conditional Variance

While the conditional mean is a good measure of the location of a conditional distribution, it does not provide information about the spread of the distribution. A common measure of the dispersion is the **conditional variance**

$$\sigma^2(x) = Var\left(y_i \mid x_i = x\right) = E\left(e_i^2 \mid x_i = x\right).$$

Generally, $\sigma^2(x)$ is a non-trivial function of x, and can take any form, subject to the restriction that it is non-negative. The **conditional standard deviation** is its square root $\sigma(x) = \sqrt{\sigma^2(x)}$.

Given the random variable x_i, the conditional variance of y_i is $\sigma_i^2 = \sigma^2(x_i)$. In the general case where $\sigma^2(x)$ depends on x we say that the error e_i is **heteroskedastic**. In contrast, when $\sigma^2(x)$ is a constant so that

$$E\left(e_i^2 \mid x_i\right) = \sigma^2 \tag{3.4}$$

we say that the error e_i is **homoskedastic**.

Some textbooks inappropriately describe heteroskedasticity as the case where "the variance of e_i varies across observation i". This concept is less helpful than defining heteroskedasticity as the dependence of the conditional variance on the observables x_i.

As an example, take the conditional wage densities displayed in Figure 3.1. The conditional standard deviation for men is 12.1 and that for women is 10.5. So while men have higher average wages, they are also somewhat more dispersed.

3.4 Linear Regression

An important special case of (3.3) is when the conditional mean function $m(x)$ is linear in x (or linear in functions of x). Notationally, it is convenient to augment the regressor vector x_i by listing the number "1" as an element. We call this the "constant" or "intercept". Equivalently, we assume that $x_{1i} = 1$, where x_{1i} is the first element of the vector x_i defined in (3.1). Thus (3.1) has been redefined as the $k \times 1$ vector

$$x_i = \begin{pmatrix} 1 \\ x_{2i} \\ \vdots \\ x_{ki} \end{pmatrix}. \tag{3.5}$$

When $m(x)$ is linear in x, we can write it as

$$m(x) = x'\beta = \beta_1 + x_{2i}\beta_2 + \cdots + x_{ki}\beta_k \tag{3.6}$$

where

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \tag{3.7}$$

is a $k \times 1$ parameter vector.

In this case (3.3) can be writen as

$$\begin{aligned} y_i &= x_i'\beta + e_i & (3.8) \\ E\left(e_i \mid x_i\right) &= 0. & (3.9) \end{aligned}$$

Equation (3.8) is called the **linear regression model,**

An important special case is **homoskedastic linear regression model**

$$\begin{aligned} y_i &= x_i'\beta + e_i \\ E\left(e_i \mid x_i\right) &= 0 \\ E\left(e_i^2 \mid x_i\right) &= \sigma^2. \end{aligned}$$

3.5 Best Linear Predictor

While the conditional mean $m(x) = E\left(y_i \mid x_i = x\right)$ is the best predictor of y_i among all functions of x_i, its functional form is typically unknown, and the linear assumption of the previous section is empirically unlikely to accurate. Instead, it is more realistic to view the linear specification (3.6) as an approximation, which we derive in this section.

In the **linear projection model** the coefficient β is defined so that the function $x_i'\beta$ is the best linear predictor of y_i. As before, by "best" we mean the predictor function with lowest mean squared error. For any $\beta \in R^k$ a linear predictor for y_i is $x_i'\beta$ with expected squared prediction error

$$\begin{aligned} S(\beta) &= E\left(y_i - x_i'\beta\right)^2 \\ &= Ey_i^2 - 2E\left(y_i x_i'\right)\beta + \beta' E\left(x_i x_i'\right)\beta. \end{aligned}$$

which is quadratic in β. The **best linear predictor** is obtained by selecting β to minimize $S(\beta)$. The first-order condition for minimization (from Section 2.7) is

$$0 = \frac{\partial}{\partial \beta} S(\beta) = -2E\left(x_i y_i\right) + 2E\left(x_i x_i'\right)\beta.$$

Solving for β we find

$$\beta = \left(E\left(x_i x_i'\right)\right)^{-1} E\left(x_i y_i\right). \tag{3.10}$$

It is worth taking the time to understand the notation involved in this expression. $E\left(x_i x_i'\right)$ is a matrix and $E\left(x_i y_i\right)$ is a vector. Therefore, alternative expressions such as $\frac{E(x_i y_i)}{E\left(x_i x_i'\right)}$ or $E\left(x_i y_i\right)\left(E\left(x_i x_i'\right)\right)^{-1}$ are incoherent and incorrect.

The vector (3.10) exists and is unique as long as the $k \times k$ matrix $E\left(x_i x_i'\right)$ is invertible. Observe that for any non-zero $\alpha \in R^k$, $\alpha' E\left(x_i x_i'\right)\alpha = E\left(\alpha' x_i\right)^2 \geq 0$ so the matrix $E\left(x_i x_i'\right)$ is by construction positive semi-definite. It is invertible if and only if it is positive definite, written $E\left(x_i x_i'\right) > 0$, which requires that for all non-zero α, $\alpha' E\left(x_i x_i'\right)\alpha = E\left(\alpha' x_i\right)^2 > 0$. Equivalently, there cannot exist a non-zero vector α such that $\alpha' x_i = 0$ identically. This occurs when redundant variables are included in x_i. In order for β to be uniquely defined, this situation must be excluded.

Given the definition of β in (3.10), $x_i'\beta$ is the best linear predictor for y_i. The **error** is

$$e_i = y_i - x_i'\beta. \tag{3.11}$$

Notice that the error from the linear prediction equation e_i is equal to the error from the regression equation when (and only when) the conditional mean is linear in x_i, otherwise they are distinct.

Rewriting, we obtain a decomposition of y_i into linear predictor and error

$$y_i = x_i'\beta + e_i. \tag{3.12}$$

This completes the derivation of the linear projection model. We now summarize the assumptions necessary for its derivation and list the implications in Theorem 3.5.1.

Assumption 3.5.1

1. *x_i contains an intercept;*
2. *$Ey_i^2 < \infty$;*
3. *$E\left(x_i' x_i\right) < \infty$;*
4. *$E\left(x_i x_i'\right)$ is invertible.*

Theorem 3.5.1 *Under Assumption 3.5.1, (3.10) and (3.11) are well defined. Furthermore,*

$$E\left(x_i e_i\right) = 0 \tag{3.13}$$

and

$$E\left(e_i\right) = 0. \tag{3.14}$$

Proof. Assumption 3.5.1.2 and 3.5.1.3 ensure that the moments in (3.10) are defined. Assumption 3.5.1.4 guarantees that the solution β exits. Using the definitions (3.11) and (3.10)

$$\begin{aligned} E\left(x_i e_i\right) &= E\left(x_i\left(y_i - x_i'\beta\right)\right) \\ &= E\left(x_i y_i\right) - E\left(x_i x_i'\right)\left(E\left(x_i x_i'\right)\right)^{-1} E\left(x_i y_i\right) \\ &= 0. \end{aligned}$$

Equation (3.14) follows from (3.13) and Assumption 3.5.1.1. ■

The two equations (3.12) and (3.13) summarize the linear projection model. Let's compare it with the linear regression model (3.8)-(3.9). Since from Theorem 3.2.1.4 we know that the regression error has the property $E\left(x_i e_i\right) = 0$, it follows that linear regression is a special case of the projection model. However, the converse is not true as the projection error does not necessarily satisfy $E\left(e_i \mid x_i\right) = 0$. For example, suppose that for $x_i \in R$ that $Ex_i^3 = 0$ and $e_i = x_i^2$. Then $Ex_i e_i^2 = Ex_i^3 = 0$ yet $E\left(e_i \mid x_i\right) = x_i^2 \neq 0$.

Since $E\left(x_i e_i\right) = 0$ we say that x_i and e_i are **orthogonal**. This means that the equation (3.12) can be alternatively interpreted as a **projection decomposition**. By definition, $x_i'\beta$ is the **projection** of y_i on x_i since the error e_i is orthogonal with x_i. Since e_i is mean zero by (3.14), the orthogonality restriction (3.13) implies that x_i and e_i are uncorrelated.

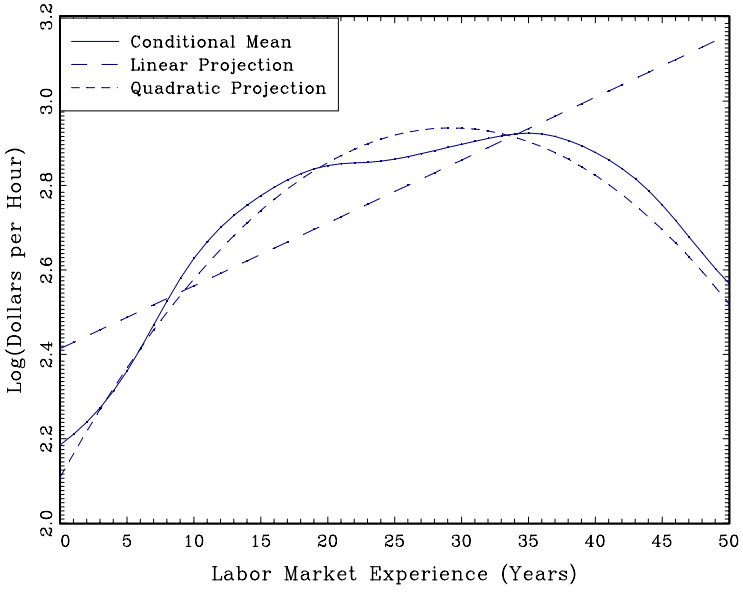

Figure 3.4: Hourly Wage as a Function of Experience

The conditions listed in Assumption 3.5.1 are weak. The finite variance Assumptions 3.5.1.2 and 3.5.1.3 are called **regularity** conditions. Assumption 3.5.1.4 is required to ensure that β is uniquely defined. Assumption 3.5.1.1 is employed to guarantee that (3.14) holds.

We have shown that under mild regularity conditions, for any pair (y_i, x_i) we can define a linear projection equation (3.12) with the properties listed in Theorem 3.5.1. No additional assumptions are required. However, it is important to not misinterpret the generality of this statement. The linear equation (3.12) is defined by projection and the associated coefficient definition (3.10). In contrast, in many economic models the parameter β may be defined within the model. In this case (3.10) may not hold and the implications of Theorem 3.5.1 may be false. These structural models require alternative estimation methods, and are discussed in Chapter 11.

Returning to the joint distribution displayed in Figure 3.3, the dashed line is the linear projection of log wages on eduction. In this example the linear projection is a close approximation to the conditional mean. In other cases the two may be quite different. Figure 3.4 displays the relationship[3] between mean log hourly wages and labor market experience. The solid line is the conditional mean, and the straight dashed line is the linear projection. In this case the linear projection is a poor approximation to the conditional mean. It over-predicts wages for young and old workers, and under-predicts for the rest. Most importantly, it misses the strong downturn in expected wages for those above 35 years work experience (equivalently, for those over 53 in age).

This defect in linear projection can be partially corrected through a careful selection of regressors. In the example just presented, we can augment the regressor vector x_i to include both *experience* and *experience*^{2}. A projection of log wages on these two variables can be called a quadratic projection, since the resulting function is quadratic in *experience*. Other than the redefinition of the regressor vector, there are no changes in our methods or analysis. In Figure 1.4 we display as well this quadratic projection. In this example it is a much better approximation to the conditional mean than the linear projection.

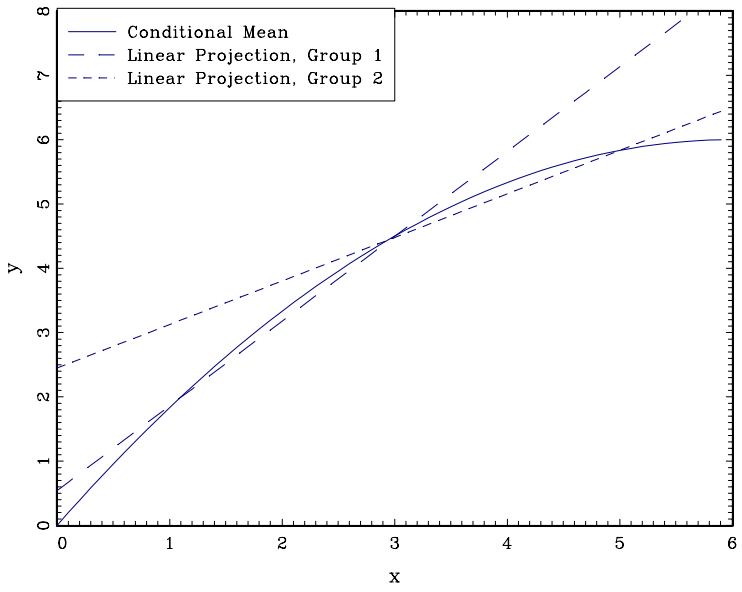

Figure 3.5: Conditional Mean and Two Linear Projections

Another defect of linear projection is that it is sensitive to the marginal distribution of the regressors when the conditional mean is non-linear. We illustrate the issue in Figure 3.5 for a

[3] In the population of Caucasian non-military male wage earners with 12 years of education.

constructed[4] joint distribution of y_i and x_i. The solid line is the non-linear conditional mean of y_i given x_i. The data are divided in two – Group 1 and Group 2 – which have different marginal distributions for the regressor x_i, and Group 1 has a lower mean value of x_i than Group 2. The separate linear projections of y_i on x_i for these two groups is displayed in the Figure with the dashed lines. These two projections are distinct approximations to the conditional mean. A defect with linear projection is that it leads to the incorrect conclusion that the effect of x_i on y_i is different for individuals in the two Groups. This conclusion is incorrect because is fact there is no difference in the conditional mean between the two groups. The apparant difference is a by-product of a linear approximation to a non-linear mean, combined with different marginal distributions for the conditioning variables.

[4]The x_i in Group 1 are $N(2,1)$ and those in Group 2 are $N(4,1)$, and the conditional distriubtion of y_i given x_i is $N(m(x_i),1)$ where $m(x) = 2x - x^2/6$.

3.6 Exercises

1. Prove parts 2, 3 and 4 of Theorem 3.2.1.

2. Suppose that Y and X only take the values 0 and 1, and have the following joint probability distribution

	$X=0$	$X=1$
$Y=0$	.1	.2
$Y=1$	.4	.3

   Find $E(Y \mid X=x)$, $E(Y^2 \mid X=x)$ and $Var(Y \mid X=x)$.

3. Suppose that y_i is discrete-valued, taking values only on the non-negative integers, and the conditional distribution of y_i given x_i is Poisson:

$$P\left(y_i=k \mid x_i=x\right)=\frac{e^{-x'\beta}\left(x'\beta\right)^j}{j!}, \qquad j=0,1,2,...$$

   Compute $E(y_i \mid x_i=x)$ and $Var(y_i \mid x_i=x)$. Does this justify a linear regression model of the form $y_i=x_i'\beta+\varepsilon_i$?

   Hint: If $P\left(Y=j\right)=\frac{e^{-\lambda}\lambda^j}{j!}$, then $EY=\lambda$ and $Var(Y)=\lambda$.

4. Let x_i and y_i have the joint density $f(x,y)=\frac{3}{2}\left(x^2+y^2\right)$ on $0\leq x\leq 1$, $0\leq y\leq 1$. Compute the coefficients of the linear projection $y_i=\beta_0+\beta_1 x_i+e_i$. Compute the conditional mean $m(x)=E\left(y_i \mid x_i=x\right)$. Are they different?

5. Take the bivariate linear projection model

$$\begin{aligned} y_i &= \beta_0+\beta_1 x_i+e_i \\ E\left(e_i\right) &= 0 \\ E\left(x_i e_i\right) &= 0 \end{aligned}$$

   Define $\mu_y=Ey_i$, $\mu_x=Ex_i$, $\sigma_x^2=Var(x_i)$, $\sigma_y^2=Var(y_i)$ and $\sigma_{xy}=Cov(x_i,y_i)$. Show that $\beta_1=\sigma_{xy}/\sigma_x^2$ and $\beta_0=\mu_y-\beta_1\mu_x$.

6. True or False. If $y_i=x_i\beta+e_i$, $x_i\in R$, and $E(e_i \mid x_i)=0$, then $E(x_i^2 e_i)=0$.

7. True or False. If $y_i=x_i'\beta+e_i$ and $E(e_i \mid x_i)=0$, then e_i is independent of x_i.

8. True or False. If $y_i=x_i'\beta+e_i$, $E(e_i \mid x_i)=0$, and $E(e_i^2 \mid x_i)=\sigma^2$, a constant, then e_i is independent of x_i.

9. True or False. If $y_i=x_i\beta+e_i$, $x_i\in R$, and $E(x_i e_i)=0$, then $E(x_i^2 e_i)=0$.

10. True or False. If $y_i=x_i'\beta+e_i$ and $E(x_i e_i)=0$, then $E(e_i \mid x_i)=0$.

11. Let X be a random variable with $\mu=EX$ and $\sigma^2=Var(X)$. Define

$$g\left(x,\mu,\sigma^2\right)=\begin{pmatrix} x-\mu \\ (x-\mu)^2-\sigma^2 \end{pmatrix}.$$

    Show that $Eg\left(X,m,s\right)=0$ if and only if $m=\mu$ and $s=\sigma^2$.

Chapter 4

Least Squares Estimation

This chapter explores estimation and inference in the linear projection model

$$\begin{aligned} y_i &= x_i'\beta + e_i && (4.1)\\ E(x_i e_i) &= 0 && (4.2)\\ \beta &= \left(E\left(x_i x_i'\right)\right)^{-1} E(x_i y_i) && (4.3) \end{aligned}$$

In Sections 4.7 and 4.8, we narrow the focus to the linear regression model, but for most of the chapter we retain the broader focus on the projection model.

4.1 Estimation

Equation (4.3) writes the projection coefficient β as an explicit function of population moments $E(x_i y_i)$ and $E(x_i x_i')$. Their moment estimators are the sample moments

$$\begin{aligned} \hat{E}(x_i y_i) &= \frac{1}{n}\sum_{i=1}^n x_i y_i \\ \hat{E}\left(x_i x_i'\right) &= \frac{1}{n}\sum_{i=1}^n x_i x_i'. \end{aligned}$$

It follows that the moment estimator of β replaces the population moments in (4.3) with the sample moments:

$$\begin{aligned} \hat{\beta} &= \left(\hat{E}\left(x_i x_i'\right)\right)^{-1} \hat{E}(x_i y_i) \\ &= \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^n x_i y_i \\ &= \left(\sum_{i=1}^n x_i x_i'\right)^{-1} \sum_{i=1}^n x_i y_i. \qquad (4.4) \end{aligned}$$

Another way to derive $\hat{\beta}$ is as follows. Observe that (4.2) can be written in the parametric form $g(\beta) = E\left(x_i\left(y_i - x_i'\beta\right)\right) = 0$. The function $g(\beta)$ can be estimated by

$$\hat{g}(\beta) = \frac{1}{n}\sum_{i=1}^n x_i\left(y_i - x_i'\beta\right).$$

This is a set of k equations which are linear in β. The estimator $\hat{\beta}$ is the value which jointly sets these equations equal to zero:

$$\begin{aligned} 0 &= \hat{g}(\hat{\beta}) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i\left(y_i - x_i'\hat{\beta}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i x_i'\hat{\beta} \end{aligned} \tag{4.5}$$

whose solution is (4.4).

To illustrate, consider the data used to generate Figure 3.3. These are white male wage earners from the March 2004 Current Population Survey, excluding military, with 10-15 years of potential work experience. This sample has 988 observations. Let y_i be log wages and x_i be an intercept and years of education. Then

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i = \begin{pmatrix} 2.95 \\ 42.40 \end{pmatrix}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' = \begin{pmatrix} 1 & 14.14 \\ 14.14 & 205.83 \end{pmatrix}.$$

Thus

$$\begin{aligned} \hat{\beta} &= \begin{pmatrix} 1 & 14.14 \\ 14.14 & 205.83 \end{pmatrix}^{-1} \begin{pmatrix} 2.95 \\ 42.40 \end{pmatrix} \\ &= \begin{pmatrix} 34.94 & -2.40 \\ -2.40 & 0.170 \end{pmatrix} \begin{pmatrix} 2.71 \\ 37.37 \end{pmatrix} \\ &= \begin{pmatrix} 1.30 \\ 0.117 \end{pmatrix}. \end{aligned}$$

We often write the estimated equation using the format

$$\widehat{\log(Wage_i)} = 1.30 + 0.117\ Education_i.$$

An interpretation of the estimated equation is that each year of education is associated with an 11.7% increase in mean wages.

4.2 Least Squares

There is another classic motivation for the estimator (4.4). Define the **sum-of-squared errors** (SSE) function

$$\begin{aligned} S_n(\beta) &= \sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2 \\ &= \sum_{i=1}^{n} y_i^2 - 2\beta'\sum_{i=1}^{n} x_i y_i + \beta'\sum_{i=1}^{n} x_i x_i'\beta. \end{aligned}$$

This is a quadratic function of β.

The **Ordinary Least Squares (OLS)** estimator is the value of β which minimizes $S_n(\beta)$. Matrix calculus (see Appendix 2.7) gives the first-order conditions for minimization:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \beta} S_n(\hat{\beta}) \\ &= -2\sum_{i=1}^{n} x_i y_i + 2\sum_{i=1}^{n} x_i x_i' \hat{\beta} \end{aligned}$$

whose solution is (4.4). Following convention we will call $\hat{\beta}$ the OLS estimator of β.

To visualize the sum-of-squared errors function, Figure 4.1 displays an example sum–of-squared errors function $S_n(\beta)$ for the case $k = 2$. Figure 4.2 displays the contour lines of the same function – horizontal slices at equally spaced heights. Since the function $S_n(\beta)$ is a quadratic function of β, the contour lines are ellipses.

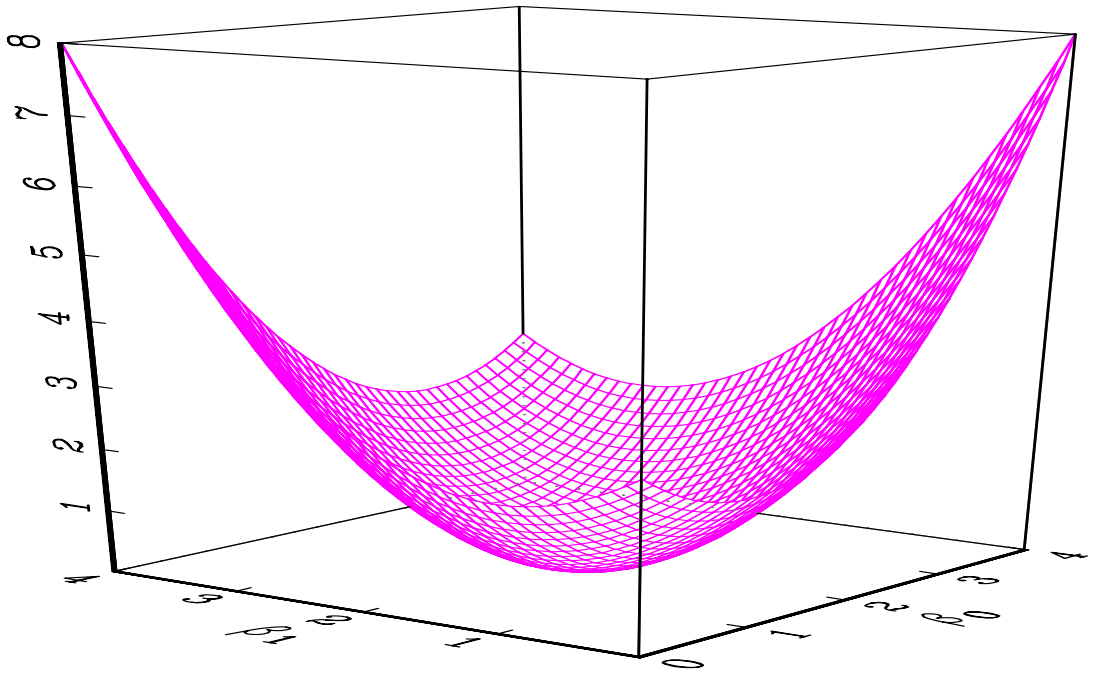

Figure 4.1: Sum-of-Squared Errors Function

As a by-product of OLS estimation, we define the **predicted value**

$$\hat{y}_i = x_i' \hat{\beta}$$

and the **residual**

$$\begin{aligned} \hat{e}_i &= y_i - \hat{y}_i \\ &= y_i - x_i' \hat{\beta}. \end{aligned}$$

Note that $y_i = \hat{y}_i + \hat{e}_i$. It is important to understand the distinction between the error e_i and the residual $\hat{e}_i$. The error is unobservable, while the residual is a by-product of estimation. These two variables are frequently mislabeled, which can cause confusion.

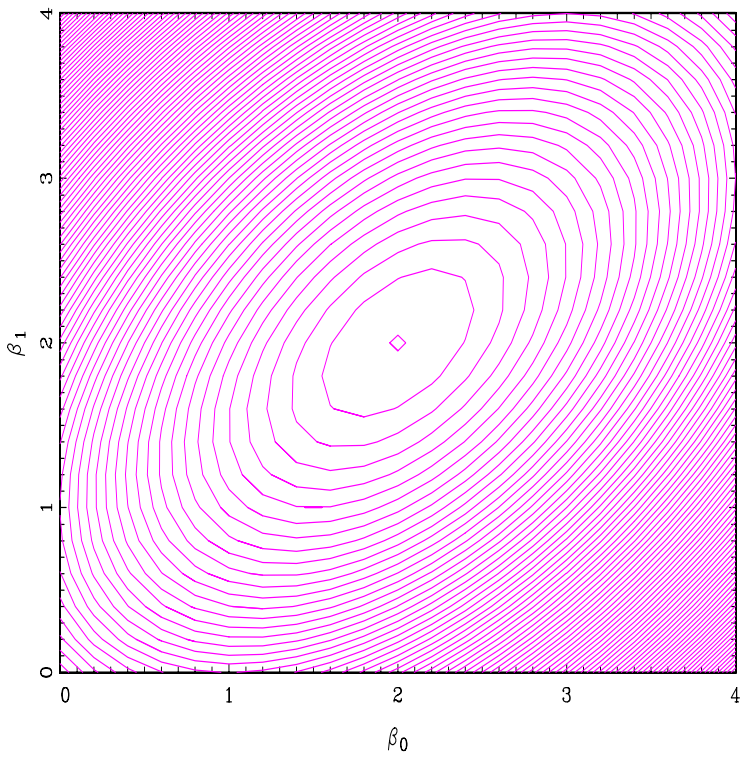

Figure 4.2: Sum-of-Squared Error Function Contours

Equation (4.5) implies that

$$\frac{1}{n}\sum_{i=1}^{n} x_i \hat{e}_i = 0.$$

Since x_i contains a constant, one implication is that

$$\frac{1}{n}\sum_{i=1}^{n} \hat{e}_i = 0.$$

Thus the residuals have a sample mean of zero and the sample correlation between the regressors and the residual is zero. These are algebraic results, and hold true for all linear regression estimates.

The error variance $\sigma^2 = Ee_i^2$ is also a parameter of interest. It measures the variation in the "unexplained" part of the regression. Its method of moments estimator is the sample average

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2. \tag{4.6}$$

An alternative estimator uses the formula

$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n} \hat{e}_i^2. \tag{4.7}$$

A justification for the latter choice will be provided in Section 4.7.

A measure of the explained variation relative to the total variation is the **coefficient of determination** or **R-squared**.

$$R^2 = \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2}$$

where

$$\hat{\sigma}_y^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \overline{y})^2$$

is the sample variance of y_i. The R^2 is frequently mislabeled as a measure of "fit". It is an inappropriate label as the value of R^2 does not help interpret the parameter estimates $\hat{\beta}$ or test statistics concerning β. Instead, it should be viewed as an estimator of the population parameter

$$\rho^2 = \frac{Var\left(x_i'\beta\right)}{Var(y_i)} = 1 - \frac{\sigma^2}{\sigma_y^2}$$

where $\sigma_y^2 = Var(y_i)$. An alternative estimator of ρ^2 proposed by Theil called "R-bar-squared" is

$$\overline{R}^2 = 1 - \frac{s^2}{\tilde{\sigma}_y^2}$$

where

$$\tilde{\sigma}_y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \overline{y})^2.$$

Theil's estimator $\overline{R}^2$ is a ratio of adjusted variance estimators, and therefore is expected to be a better estimator of ρ^2 than the unadjusted estimator R^2.

4.3 Normal Regression Model

Another motivation for the least-squares estimator can be obtained from the normal regression model. This is the linear regression model with the additional assumption that the error e_i is independent of x_i and has the distribution $N(0, \sigma^2)$. This is a parametric model, where likelihood methods can be used for estimation, testing, and distribution theory.

The log-likelihood function for the normal regression model is

$$\begin{aligned} L_n(\beta, \sigma^2) &= \sum_{i=1}^n \log\left(\frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - x_i'\beta\right)^2\right)\right) \\ &= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} S_n(\beta) \end{aligned}$$

The MLE $(\hat{\beta}, \hat{\sigma}^2)$ maximize $L_n(\beta, \sigma^2)$. Since $L_n(\beta, \sigma^2)$ is a function of β only through the sum of squared errors $S_n(\beta)$, maximizing the likelihood is identical to minimizing $S_n(\beta)$. Hence the MLE for β equals the OLS estimator.

Plugging $\hat{\beta}$ into the log-likelihood we obtain

$$L_n(\hat{\beta}, \sigma^2) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^n \hat{e}_i^2$$

Maximization with respect to σ^2 yields the first-order condition

$$\frac{\partial}{\partial \sigma^2} L_n(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\left(\hat{\sigma}^2\right)^2} \sum_{i=1}^{n} \hat{e}_i^2 = 0.$$

Solving for $\hat{\sigma}^2$ yields the method of moments estimator (4.6). Thus the MLE $(\hat{\beta}, \hat{\sigma}^2)$ for the normal regression model are identical to the method of moment estimators. Due to this equivalence, the OLS estimator $\hat{\beta}$ is frequently referred to as the Gaussian MLE.

4.4 Model in Matrix Notation

For many purposes, including computation, it is convenient to write the model and statistics in matrix notation. We define

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix}, \qquad e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

Observe that Y and e are $n \times 1$ vectors, and X is an $n \times k$ matrix.

The linear equation (3.12) is a system of n equations, one for each observation. We can stack these n equations together as

$$\begin{aligned} y_1 &= x_1'\beta + e_1 \\ y_2 &= x_2'\beta + e_2 \\ &\vdots \\ y_n &= x_n'\beta + e_n. \end{aligned}$$

or equivalently

$$Y = X\beta + e.$$

Sample sums can also be written in matrix notation. For example

$$\begin{aligned} \sum_{i=1}^{n} x_i x_i' &= X'X \\ \sum_{i=1}^{n} x_i y_i &= X'Y. \end{aligned}$$

Thus the estimator (4.4), residual vector, and sample error variance can be written as

$$\begin{aligned} \hat{\beta} &= \left(X'X\right)^{-1}\left(X'Y\right) \\ \hat{e} &= Y - X\hat{\beta} \\ \hat{\sigma}^2 &= n^{-1}\hat{e}'\hat{e}. \end{aligned}$$

A useful result is obtained by inserting $Y = X\beta + e$ into the formula for $\hat{\beta}$ to obtain

$$\begin{aligned} \hat{\beta} &= \left(X'X\right)^{-1}\left(X'\left(X\beta + e\right)\right) \\ &= \left(X'X\right)^{-1}X'X\beta + \left(X'X\right)^{-1}\left(X'e\right) \\ &= \beta + \left(X'X\right)^{-1}X'e. \end{aligned} \tag{4.8}$$

4.5 Projection Matrices

Define the matrices

$$P = X\left(X'X\right)^{-1}X'$$

and

$$\begin{aligned} M &= I_n - X\left(X'X\right)^{-1}X' \\ &= I_n - P \end{aligned}$$

where I_n is the $n \times n$ identity matrix. They are called **projection matrices** due to the property that for any matrix Z which can be written as $Z = X\Gamma$ for some matrix Γ, (we say that Z lies in the **range space** of X) then

$$PZ = PX\Gamma = X\left(X'X\right)^{-1}X'X\Gamma = X\Gamma = Z$$

and

$$MZ = \left(I_n - P\right)Z = Z - PZ = Z - Z = 0.$$

As an important example of this property, partition the matrix X into two matrices X_1 and X_2, so that

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}.$$

Then $PX_1 = X_1$ and $MX_1 = 0$. It follows that $MX = 0$ and $MP = 0$, so M and P are orthogonal.

The matrices P and M are symmetric and **idempotent**[1]. To see that P is symmetric,

$$\begin{aligned} P' &= \left(X\left(X'X\right)^{-1}X'\right)' \\ &= \left(X'\right)'\left(\left(X'X\right)^{-1}\right)'\left(X\right)' \\ &= X\left(\left(X'X\right)'\right)^{-1}X' \\ &= X\left(\left(X\right)'\left(X'\right)'\right)^{-1}X' \\ &= P. \end{aligned}$$

To establish that it is idempotent,

$$\begin{aligned} PP &= \left(X\left(X'X\right)^{-1}X'\right)\left(X\left(X'X\right)^{-1}X'\right) \\ &= X\left(X'X\right)^{-1}X'X\left(X'X\right)^{-1}X' \\ &= X\left(X'X\right)^{-1}X' \\ &= P, \end{aligned}$$

and

$$\begin{aligned} MM &= M\left(I_n - P\right) \\ &= M - MP \\ &= M \end{aligned}$$

[1] A matrix A is **idempotent** if $AA = A$.

since $MP = 0$.

Another useful property is that

$$\begin{aligned} \operatorname{tr} P &= k \quad (4.9) \\ \operatorname{tr} M &= n - k \quad (4.10) \end{aligned}$$

where the trace operator

$$\operatorname{tr} A = \sum_{j=1}^{r} a_{jj}$$

is the sum of the diagonal elements of the matrix A.

To show (4.9) and (4.10),

$$\begin{aligned} \operatorname{tr} P &= \operatorname{tr}\left(X\left(X'X\right)^{-1}X'\right) \\ &= \operatorname{tr}\left(\left(X'X\right)^{-1}X'X\right) \\ &= \operatorname{tr}\left(I_k\right) \\ &= k, \end{aligned}$$

and

$$\operatorname{tr} M = \operatorname{tr}\left(I_n - P\right) = \operatorname{tr}\left(I_n\right) - \operatorname{tr}\left(P\right) = n - k.$$

Given the definitions of P and M, observe that

$$\hat{Y} = X\hat{\beta} = X\left(X'X\right)^{-1}X'Y = PY$$

and

$$\hat{e} = Y - X\hat{\beta} = Y - PY = MY. \quad (4.11)$$

Furthermore, since $Y = X\beta + e$ and $MX = 0$, then

$$\hat{e} = M\left(X\beta + e\right) = Me. \quad (4.12)$$

Another way of writing (4.11) is

$$Y = \left(P + M\right)Y = PY + MY = \hat{Y} + \hat{e}.$$

This decomposition is **orthogonal**, that is

$$\hat{Y}'\hat{e} = \left(PY\right)'\left(MY\right) = Y'PMY = 0.$$

4.6 Residual Regression

Partition

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$$

and

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Then the regression model can be rewritten as

$$Y = X_1\beta_1 + X_2\beta_2 + e. \quad (4.13)$$

Observe that the OLS estimator of $\beta = (\beta_1', \beta_2')'$ can be obtained by regression of Y on $X = [X_1\ X_2]$. OLS estimation can be written as

$$Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e}. \tag{4.14}$$

Suppose that we are primarily interested in β_2, not in β_1, so are only interested in obtaining the OLS sub-component $\hat{\beta}_2$. In this section we derive an alternative expression for $\hat{\beta}_2$ which does not involve estimation of the full model.

Define

$$M_1 = I_n - X_1\left(X_1'X_1\right)^{-1}X_1'.$$

Recalling the definition $M = I - X\left(X'X\right)^{-1}X'$, observe that $X_1'M_1 = 0$ and thus

$$M_1M = M - X_1\left(X_1'X_1\right)^{-1}X_1'M = M$$

It follows that

$$M_1\hat{e} = M_1Me = Me = \hat{e}.$$

Using this result, if we premultiply (4.14) by M_1 we obtain

$$\begin{aligned} M_1Y &= M_1X_1\hat{\beta}_1 + M_1X_2\hat{\beta}_2 + M_1\hat{e} \\ &= M_1X_2\hat{\beta}_2 + \hat{e} \end{aligned} \tag{4.15}$$

the second equality since $M_1X_1 = 0$. Premultiplying by X_2' and recalling that $X_2'\hat{e} = 0$, we obtain

$$X_2'M_1Y = X_2'M_1X_2\hat{\beta}_2 + X_2'\hat{e} = X_2'M_1X_2\hat{\beta}_2.$$

Solving,

$$\hat{\beta}_2 = \left(X_2'M_1X_2\right)^{-1}\left(X_2'M_1Y\right)$$

an alternative expression for $\hat{\beta}_2$.

Now, define

$$\tilde{X}_2 = M_1X_2 \tag{4.16}$$

$$\tilde{Y} = M_1Y, \tag{4.17}$$

the least-squares residuals from the regression of X_2 and Y, respectively, on the matrix X_1 only. Since the matrix M_1 is idempotent, $M_1 = M_1M_1$ and thus

$$\begin{aligned} \hat{\beta}_2 &= \left(X_2'M_1X_2\right)^{-1}\left(X_2'M_1Y\right) \\ &= \left(X_2'M_1M_1X_2\right)^{-1}\left(X_2'M_1M_1Y\right) \\ &= \left(\tilde{X}_2'\tilde{X}_2\right)^{-1}\left(\tilde{X}_2'\tilde{Y}\right) \end{aligned}$$

This shows that $\hat{\beta}_2$ can be calculated by the OLS regression of $\tilde{Y}$ on $\tilde{X}_2$. This technique is called **residual regression**.

Furthermore, using the definitions (4.16) and (4.17), expression (4.15) can be equivalently written as

$$\tilde{Y} = \tilde{X}_2\hat{\beta}_2 + \hat{e}.$$

Since $\hat{\beta}_2$ is precisely the OLS coefficient from a regression of $\tilde{Y}$ on $\tilde{X}_2$, this shows that the residual from this regression is $\hat{e}$, numerically the same residual as from the joint regression (4.14). We have proven the following theorem.

Theorem 4.6.1 *(Frisch-Waugh-Lovell). In the model (4.13), the OLS estimator of β_2 and the OLS residuals $\hat{e}$ may be equivalently computed by either the OLS regression (4.14) or via the following algorithm:*

1. *Regress Y on X_1, obtain residuals $\tilde{Y}$;*
2. *Regress X_2 on X_1, obtain residuals $\tilde{X}_2$;*
3. *Regress $\tilde{Y}$ on $\tilde{X}_2$, obtain OLS estimates $\hat{\beta}_2$ and residuals $\hat{e}$.*

In some contexts, the FWL theorem can be used to speed computation, but in most cases there is little computational advantage to using the two-step algorithm. Rather, the primary use is theoretical.

A common application of the FWL theorem, which you may have seen in an introductory econometrics course, is the demeaning formula for regression. Partition $X = [X_1 \ X_2]$ where $X_1 = \iota$ is a vector of ones, and X_2 is the vector of observed regressors. In this case,

$$M_1 = I - \iota \left(\iota'\iota\right)^{-1} \iota'.$$

Observe that

$$\begin{aligned} \tilde{X}_2 &= M_1 X_2 = X_2 - \iota \left(\iota'\iota\right)^{-1} \iota' X_2 \\ &= X_2 - \overline{X}_2 \end{aligned}$$

and

$$\begin{aligned} \tilde{Y} &= M_1 Y = Y - \iota \left(\iota'\iota\right)^{-1} \iota' Y \\ &= Y - \overline{Y}, \end{aligned}$$

which are "demeaned". The FWL theorem says that $\hat{\beta}_2$ is the OLS estimate from a regression of $\tilde{Y}$ on $\tilde{X}_2$, or $y_i - \overline{y}$ on $x_{2i} - \overline{x}_2$:

$$\hat{\beta}_2 = \left(\sum_{i=1}^{n} \left(x_{2i} - \overline{x}_2\right) \left(x_{2i} - \overline{x}_2\right)' \right)^{-1} \left(\sum_{i=1}^{n} \left(x_{2i} - \overline{x}_2\right) \left(y_i - \overline{y}\right) \right).$$

Thus the OLS estimator for the slope coefficients is a regression with demeaned data.

4.7 Bias and Variance

In this and the following section we consider the special case of the linear regression model (3.8)-(3.9). In this section we derive the small sample conditional mean and variance of the OLS estimator.

By the independence of the observations and (3.9), observe that

$$E\left(e \mid X\right) = \begin{pmatrix} \vdots \\ E\left(e_i \mid X\right) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ E\left(e_i \mid x_i\right) \\ \vdots \end{pmatrix} = 0. \tag{4.18}$$

Using (4.8), the properties of conditional expectations, and (4.18), we can calculate

$$\begin{aligned} E\left(\hat{\beta}-\beta \mid X\right) &= E\left(\left(X'X\right)^{-1}X'e \mid X\right) \\ &= \left(X'X\right)^{-1}X'E\left(e \mid X\right) \\ &= 0. \end{aligned}$$

We have shown that

$$E\left(\hat{\beta} \mid X\right) = \beta \tag{4.19}$$

which implies

$$E\left(\hat{\beta}\right) = \beta$$

and thus the OLS estimator $\hat{\beta}$ is unbiased for β.

Next, for any random vector Z define the covariance materix

$$\begin{aligned} Var(Z) &= E\left(Z-EZ\right)\left(Z-EZ\right)' \\ &= EZZ'-\left(EZ\right)\left(EZ\right)'. \end{aligned}$$

Then given (4.19) we see that

$$\begin{aligned} Var\left(\hat{\beta} \mid X\right) &= E\left(\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)' \mid X\right) \\ &= \left(X'X\right)^{-1}X'DX\left(X'X\right)^{-1} \end{aligned}$$

where

$$D = E\left(ee' \mid X\right).$$

The i'th diagonal element of D is

$$E\left(e_i^2 \mid X\right) = E\left(e_i^2 \mid x_i\right) = \sigma_i^2$$

while the ij'th off-diagonal element of D is

$$E\left(e_ie_j \mid X\right) = E\left(e_i \mid x_i\right)E\left(e_j \mid x_j\right) = 0.$$

Thus D is a diagonal matrix with i'th diagonal element σ_i^2:

$$D = diag\{\sigma_1^2, ..., \sigma_n^2\} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}. \tag{4.20}$$

In the special case of the linear homoskedastic regression model, $\sigma_i^2 = \sigma^2$ and we have the simplifications $D = I_n\sigma^2$, $X'DX = X'X\sigma^2$, and

$$Var\left(\hat{\beta} \mid X\right) = \left(X'X\right)^{-1}\sigma^2.$$

We now calculate the finite sample bias of the method of moments estimator $\hat{\sigma}^2$ for σ^2, under the additional assumption of conditional homoskedasticity $E\left(e_i^2 \mid x_i\right) = \sigma^2$. From (4.12), the properties of projection matrices, and the trace operator observe that.

$$\hat{\sigma}^2 = \frac{1}{n}\hat{e}'\hat{e} = \frac{1}{n}e'MMe = \frac{1}{n}e'Me = \frac{1}{n}\operatorname{tr}\left(e'Me\right) = \frac{1}{n}\operatorname{tr}\left(Mee'\right)$$

Then

$$\begin{aligned} E\left(\hat{\sigma}^2 \mid X\right) &= \frac{1}{n}\operatorname{tr}\left[E\left(Mee' \mid X\right)\right] \\ &= \frac{1}{n}\operatorname{tr}\left[ME\left(ee' \mid X\right)\right] \\ &= \frac{1}{n}\operatorname{tr}\left[M\sigma^2\right] \\ &= \sigma^2\frac{n-k}{n}, \end{aligned}$$

the final equality by (4.10). Thus $\hat{\sigma}^2$ is biased towards zero. As an alternative, the estimator s^2 defined (4.7) is unbiased for σ^2 by this calculation. This is the justification for the common preference of s^2 over $\hat{\sigma}^2$ in empirical practice. It is important to remember, however, that this estimator is only unbiased in the special case of the homoskedastic linear regression model. It is not unbiased in the absence of homoskedasticity, or in the projection model.

4.8 Gauss-Markov Theorem

In this section we restrict attention to the homoskedastic linear regression model, which is (3.8)-(3.9) plus $E\left(e_i^2 \mid x_i\right) = \sigma^2$. Now consider the class of estimators of β which are linear functions of the vector Y, and thus can be written as

$$\tilde{\beta} = A'Y$$

where A is an $n \times k$ function of X. The least-squares estimator is the special case obtained by setting $A = X(X'X)^{-1}$. What is the best choice of A? The Gauss-Markov theorem, which we now present, says that the least-squares estimator is the best choice, as it yields the smallest variance among all unbiased linear estimators.

By a calculation similar to those of the previous section,

$$E\left(\tilde{\beta} \mid X\right) = A'X\beta,$$

so $\tilde{\beta}$ is unbiased if (and only if) $A'X = I_k$. In this case, we can write

$$\tilde{\beta}_L = A'Y = A\left(X\beta + e\right) = \beta + Ae.$$

Thus since $Var\left(e \mid X\right) = I_n\sigma^2$ under homoskedasticity,

$$Var\left(\tilde{\beta} \mid X\right) = A'Var\left(e \mid X\right)A = A'A\sigma^2.$$

The "best" linear estimator is obtained by finding the matrix A for which this variance is the smallest in the positive definite sense. The following result, known as the Gauss-Markov theorem, is a famous statement of the solution.

Theorem 4.8.1 ***Gauss-Markov****. In the homoskedastic linear regression model, the best (minimum-variance) unbiased linear estimator is OLS.*

Proof. Let A be any $n \times k$ function of X such that $A'X = I_k$. The variance of the least-squares estimator is $(X'X)^{-1}\sigma^2$ and that of $A'Y$ is $A'A\sigma^2$. It is sufficient to show that the difference $A'A - (X'X)^{-1}$ is positive semi-definite. Set $C = A - X(X'X)^{-1}$. Note that $X'C = 0$. Then we calculate that

$$\begin{aligned} A'A - (X'X)^{-1} &= \left(C + X(X'X)^{-1}\right)'\left(C + X(X'X)^{-1*}\right) - (X'X)^{-1} \\ &= C'C + C'X(X'X)^{-1} + (X'X)^{-1}X'C + (X'X)^{-1}X'X(X'X)^{-1} - (X'X)^{-1} \\ &= C'C \end{aligned}$$

The matrix $C'C$ is positive semi-definite (see Appendix 2.5) as required. ■

The Gauss-Markov theorem is an efficiency justification for the least-squares estimator, but it is quite limited in scope. Not only has the class of models has been restricted to homoskedastic linear regressions, the class of potential estimators has been restricted to linear unbiased estimators. This latter restriction is particularly unsatisfactory, as the theorem leaves open the possibility that a non-linear or biased estimator could have lower mean squared error than the least-squares estimator.

4.9 Semiparametric Efficiency

In the previous section we presented the Gauss-Markov theorem as a limited efficiency justification for the least-squares estimator. A broader justification is provided in Chamberlain (1987), who established that in the projection model the OLS estimator has the smallest asymptotic mean-squared error among feasible estimators. This property is called **semiparametric efficiency**, and is a strong justification for the least-squares estimator. We discuss the intuition behind his result in this section.

Suppose that the joint distribution of (y_i, x_i) is discrete. That is, for finite r,

$$P\left(y_i = \tau_j, \quad x_i = \xi_j\right) = \pi_j, \qquad j = 1, ..., r$$

for some constant vectors τ_j, ξ_j, and π_j. Assume that the τ_j and ξ_j are known, but the π_j are unknown. (We know the values y_i and x_i can take, but we don't know the probabilities.)

In this discrete setting, the definition (4.3) can be rewritten as

$$\beta = \left(\sum_{j=1}^{r} \pi_j \xi_j \xi_j'\right)^{-1} \left(\sum_{j=1}^{r} \pi_j \xi_j \tau_j\right) \tag{4.21}$$

Thus β is a function of $(\pi_1, ..., \pi_r)$.

As the data are multinomial, the maximum likelihood estimator (MLE) is

$$\hat{\pi}_j = \frac{1}{n}\sum_{i=1}^{n} 1(y_i = \tau_j)\, 1\left(x_i = \xi_j\right)$$

for $j = 1, ..., r$, where $1(\cdot)$ is the indicator function. That is, $\hat{\pi}_j$ is the percentage of the observations which fall in each category. The MLE $\hat{\beta}_{mle}$ for β is then the analog of (4.21) with the parameters π_j replaced by the estimates $\hat{\pi}_j$:

$$\hat{\beta}_{mle} = \left(\sum_{j=1}^{r} \hat{\pi}_j \xi_j \xi_j'\right)^{-1} \left(\sum_{j=1}^{r} \hat{\pi}_j \xi_j \tau_j\right).$$

Substituting in the expressions for $\hat{\pi}_j$,

$$\begin{aligned}\sum_{j=1}^{r}\hat{\pi}_j\xi_j\xi_j' &= \sum_{j=1}^{r}\frac{1}{n}\sum_{i=1}^{n}1\left(y_i=\tau_j\right)1\left(x_i=\xi_j\right)\xi_j\xi_j' \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{r}1\left(y_i=\tau_j\right)1\left(x_i=\xi_j\right)x_ix_i' \\ &= \frac{1}{n}\sum_{i=1}^{n}x_ix_i'\end{aligned}$$

and

$$\begin{aligned}\sum_{j=1}^{r}\hat{\pi}_j\xi_j\tau_j &= \sum_{j=1}^{r}\frac{1}{n}\sum_{i=1}^{n}1\left(y_i=\tau_j\right)1\left(x_i=\xi_j\right)\xi_j\tau_j \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{r}1\left(y_i=\tau_j\right)1\left(x_i=\xi_j\right)x_iy_i \\ &= \frac{1}{n}\sum_{i=1}^{n}x_iy_i\end{aligned}$$

Thus

$$\hat{\beta}_{mle}=\left(\frac{1}{n}\sum_{i=1}^{n}x_ix_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}x_iy_i\right)=\hat{\beta}_{ols}$$

In other words, if the data have a discrete distribution, the maximum likelihood estimator is identical to the OLS estimator. Since this is a regular parametric model the MLE is asymptotically efficient (see Appendix A.9), and thus so is the OLS estimator.

Chamberlain (1987) extends this argument to the case of continuously-distributed data. He observes that the above argument holds for all multinomial distributions, and any continuous distribution can be arbitrarily well approximated by a multinomial distribution. He proves that generically the OLS estimator (4.4) is an asymptotically efficient estimator for the parameter β defined in (3.10) for the class of models satisfying Assumption 3.5.1.

4.10 Omitted Variables

Let the regressors be partitioned as

$$x_i=\begin{pmatrix}x_{1i}\\x_{2i}\end{pmatrix}.$$

Suppose we are interested in the coefficient on x_{1i} alone in the regression of y_i on the full set x_i. We can write the model as

$$\begin{aligned}y_i &= x_{1i}'\beta_1+x_{2i}'\beta_2+e_i \\ E\left(x_ie_i\right) &= 0\end{aligned}\tag{4.22}$$

where the parameter of interest is β_1.

Now suppose that instead of estimating equation (4.22) by least-squares, we regress y_i on x_{1i} only. This is estimation of the equation

$$\begin{aligned} y_i &= x_{1i}'\gamma_1 + u_i \\ E(x_{1i}u_i) &= 0 \end{aligned} \tag{4.23}$$

Notice that we have written the coefficient on x_{1i} as γ_1 rather than β_1, and the error as u_i rather than e_i. This is because the model being estimated is different than (4.22). Goldberger (1991) calls (4.22) the **long regression** and (4.23) the **short regression** to emphasize the distinction.

Typically, $\beta_1 \neq \gamma_1$, except in special cases. To see this, we calculate

$$\begin{aligned} \gamma_1 &= \left(E\left(x_{1i}x_{1i}'\right)\right)^{-1} E\left(x_{1i}y_i\right) \\ &= \left(E\left(x_{1i}x_{1i}'\right)\right)^{-1} E\left(x_{1i}\left(x_{1i}'\beta_1 + x_{2i}'\beta_2 + e_i\right)\right) \\ &= \beta_1 + \left(E\left(x_{1i}x_{1i}'\right)\right)^{-1} E\left(x_{1i}x_{2i}'\right)\beta_2 \\ &= \beta_1 + \Gamma\beta_2 \end{aligned}$$

where

$$\Gamma = \left(E\left(x_{1i}x_{1i}'\right)\right)^{-1} E\left(x_{1i}x_{2i}'\right)$$

is the coefficient from a regression of x_{2i} on x_{1i}.

Observe that $\gamma_1 \neq \beta_1$ unless $\Gamma = 0$ or $\beta_2 = 0$. Thus the short and long regressions have the same coefficient on x_{1i} only under one of two conditions. First, the regression of x_{2i} on x_{1i} yields a set of zero coefficients (they are uncorrelated), or second, the coefficient on x_{2i} in (4.22) is zero. In general, least-squares estimation of (4.23) is an estimate of $\gamma_1 = \beta_1 + \Gamma\beta_2$ rather than β_1. The difference $\Gamma\beta_2$ is known as **omitted variable bias**. It is the consequence of omission of a relevant correlated variable.

To avoid omitted variables bias the standard advice is to include potentially relevant variables in the estimated model. By construction, the general model will be free of the omitted variables problem. Typically there are limits, as many desired variables are not available in a given dataset. In this case, the possibility of omitted variables bias should be acknowledged and discussed in the course of an empirical investigation.

4.11 Multicollinearity

If $rank(X'X) < k+1$, then $\hat{\beta}$ is not defined. This is called **strict multicollinearity**. This happens when the columns of X are linearly dependent, i.e., there is some α such that $X\alpha = 0$. Most commonly, this arises when sets of regressors are included which are identically related. For example, if X includes both the logs of two prices and the log of the relative prices, $\log(p_1)$, $\log(p_2)$ and $\log(p_1/p_2)$. When this happens, the applied researcher quickly discovers the error as the statistical software will be unable to construct $(X'X)^{-1}$. Since the error is discovered quickly, this is rarely a *problem* for applied econometric practice.

The more relevant issue is **near multicollinearity**, which is often called "multicollinearity" for brevity. This is the situation when the $X'X$ matrix is *near* singular, when the columns of X are *close* to linearly dependent. This definition is not precise, because we have not said what it means for a matrix to be "near singular". This is one difficulty with the definition and interpretation of multicollinearity.

One implication of near singularity of matrices is that the numerical reliability of the calculations is reduced. In extreme cases it is possible that the reported calculations will be in error due to floating-point calculation difficulties.

A more relevant implication of near multicollinearity is that individual coefficient estimates will be imprecise. We can see this most simply in a homoskedastic linear regression model with two regressors

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + e_i,$$

and

$$\frac{1}{n}X'X = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

In this case

$$Var\left(\hat{\beta} \mid X\right) = \frac{\sigma^2}{n}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} = \frac{\sigma^2}{n\left(1-\rho^2\right)}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

The correlation ρ indexes collinearity, since as ρ approaches 1 the matrix becomes singular. We can see the effect of collinearity on precision by observing that the asymptotic variance of a coefficient estimate $\sigma^2\left(1-\rho^2\right)^{-1}$ approaches infinity as ρ approaches 1. Thus the more "collinear" are the regressors, the worse the precision of the individual coefficient estimates.

What is happening is that when the regressors are highly dependent, it is statistically difficult to disentangle the impact of β_1 from that of β_2. As a consequence, the precision of individual estimates are reduced.

4.12 Influential Observations

The i'th observation is **influential** on the least-squares estimate if the deletion of the observation from the sample results in a meaningful change in $\hat{\beta}$. To investigate the possibility of influential observations, define the leave-one-out least-squares estimator of β, that is, the OLS estimator based on the sample *excluding* the i'th observation. This equals

$$\hat{\beta}_{(-i)} = \left(X'_{(-i)}X_{(-i)}\right)^{-1} X_{(-i)}Y_{(-i)} \tag{4.24}$$

where $X_{(-i)}$ and $Y_{(-i)}$ are the data matrices omitting the i'th row. A convenient alternative expression is

$$\hat{\beta}_{(-i)} = \hat{\beta} - (1-h_i)^{-1}\left(X'X\right)^{-1} x_i\hat{e}_i \tag{4.25}$$

where

$$h_i = x_i'\left(X'X\right)^{-1} x_i$$

is the i'th diagonal element of the projection matrix $X\left(X'X\right)^{-1}X'$. We derive expression (4.25) below.

We can also define the leave-one-out residual

$$\hat{e}_{i,-i} = y_i - x_i'\hat{\beta}_{(-i)} = (1-h_i)^{-1}\hat{e}_i. \tag{4.26}$$

A simple comparison yields that

$$\hat{e}_i - \hat{e}_{i,-i} = (1-h_i)^{-1} h_i\hat{e}_i. \tag{4.27}$$

As we can see, the change in the coefficient estimate by deletion of the i'th observation depends critically on the magnitude of h_i. The h_i take values in $[0,1]$ and sum to k. If the i'th observation

has a large value of h_i, then this observation is a **leverage point** and has the potential to be an influential observation. Investigations into the presence of influential observations can plot the values of (4.27), which is considerably more informative than plots of the uncorrected residuals $\hat{e}_i$.

We now derive equation (4.25). The key is equation (2.1) in Section 2.4 which states that

$$(A + BCD)^{-1} = A^{-1} - A^{-1}BC\left(C + CDA^{-1}BC\right)CDA^{-1}.$$

This implies

$$\left(X'X - x_i x_i'\right)^{-1} = \left(X'X\right)^{-1} + \left(X'X\right)^{-1} x_i \left(1 - h_i\right)^{-1} x_i' \left(X'X\right)^{-1}$$

and thus

$$\begin{aligned}
\hat{\beta}_{(-i)} &= \left(X'X - x_i x_i'\right)^{-1}\left(X'Y - x_i y_i\right) \\
&= \left(X'X\right)^{-1}\left(X'Y - x_i y_i\right) + \left(1 - h_i\right)^{-1}\left(X'X\right)^{-1} x_i x_i' \left(X'X\right)^{-1}\left(X'Y - x_i y_i\right) \\
&= \hat{\beta} - \left(1 - h_i\right)^{-1}\left(X'X\right)^{-1} x_i \hat{e}_i.
\end{aligned}$$

4.13 Exercises

1. Let X be a random variable with $\mu = EX$ and $\sigma^2 = Var(X)$. Define

$$g\left(x, \mu, \sigma^2\right) = \begin{pmatrix} x - \mu \\ (x-\mu)^2 - \sigma^2 \end{pmatrix}.$$

Let $(\hat{\mu}, \hat{\sigma}^2)$ be the values such that $\overline{g}_n(\hat{\mu}, \hat{\sigma}^2) = 0$ where $\overline{g}_n(m, s) = n^{-1} \sum_{i=1}^n g\left(X_i, \mu, \sigma^2\right)$. Show that $\hat{\mu}$ and $\hat{\sigma}^2$ are the sample mean and variance.

2. Consider the OLS regression of the $n \times 1$ vector y on the $n \times k$ matrix X. Consider an alternative set of regressors $Z = XC$, where C is a $k \times k$ non-singular matrix. Thus, each column of Z is a mixture of some of the columns of X. Compare the OLS estimates and residuals from the regression of Y on X to the OLS estimates from the regression of Y on Z.

3. Let $\hat{e}$ be the OLS residual from a regression of Y on $X = [X_1\ X_2]$. Find $X_2'\hat{e}$.

4. Let $\hat{e}$ be the OLS residual from a regression of Y on X. Find the OLS coefficient estimate from a regression of $\hat{e}$ on X.

5. Let $\hat{y} = X(X'X)^{-1}X'y$. Find the OLS coefficient estimate from a regression of $\hat{y}$ on X.

6. Prove that R^2 is the square of the simple correlation between y and $\hat{y}$.

7. Explain the difference between $\frac{1}{n}\sum_{i=1}^n x_i x_i'$ and $E\left(x_i x_i'\right)$.

8. Let $\hat{\beta}_n = \left(X_n'X_n\right)^{-1} X_n'Y_n$ denote the OLS estimate when Y_n is $n \times 1$ and X_n is $n \times k$. A new observation (y_{n+1}, x_{n+1}) becomes available. Prove that the OLS estimate computed using this additional observation is

$$\hat{\beta}_{n+1} = \hat{\beta}_n + \frac{1}{1 + x_{n+1}'\left(X_n'X_n\right)^{-1} x_{n+1}} \left(X_n'X_n\right)^{-1} x_{n+1}\left(y_{n+1} - x_{n+1}'\hat{\beta}_n\right).$$

9. True or False. If $y_i = x_i\beta + e_i$, $x_i \in R$, $E(e_i \mid x_i) = 0$, and $\hat{e}_i$ is the OLS residual from the regression of y_i on x_i, then $\sum_{i=1}^n x_i^2 \hat{e}_i = 0$.

10. A dummy variable takes on only the values 0 and 1. It is used for categorical data, such as an individual's gender. Let D_1 and D_2 be vectors of 1's and 0's, with the $i'th$ element of D_1 equaling 1 and that of D_2 equaling 0 if the person is a man, and the reverse if the person is a woman. Suppose that there are n_1 men and n_2 women in the sample. Consider the three regressions

$$Y = \mu + D_1\alpha_1 + D_2\alpha_2 + e \tag{4.28}$$
$$Y = D_1\alpha_1 + D_2\alpha_2 + e \tag{4.29}$$
$$Y = \mu + D_1\phi + e \tag{4.30}$$

   (a) Can all three regressions (4.28), (4.29), and (4.30) be estimated by OLS? Explain if not.

   (b) Compare regressions (4.29) and (4.30). Is one more general than the other? Explain the relationship between the parameters in (4.29) and (4.30).

(c) Compute $\iota' D_1$ and $\iota' D_2$, where ι is an $n \times 1$ is a vector of ones.

(d) Letting $\alpha = (\alpha_1' \ \alpha_2')'$, write equation (4.29) as $Y = X\alpha + e$. Consider the assumption $E(x_i e_i) = 0$. Is there any content to this assumption in this setting?

11. Let D_1 and D_2 be defined as in the previous exercise.

(a) In the OLS regression

$$Y = D_1\hat{\gamma}_1 + D_2\hat{\gamma}_2 + \hat{u},$$

show that $\hat{\gamma}_1$ is sample mean of the dependent variable among the men of the sample ($\overline{Y}_1$), and that $\hat{\gamma}_2$ is the sample mean among the women ($\overline{Y}_2$).

(b) Describe in words the transformations

$$\begin{aligned} Y^* &= Y - D_1\overline{Y}_1 + D_2\overline{Y}_2 \\ X^* &= X - D_1\overline{X}_1 + D_2\overline{X}_2. \end{aligned}$$

(c) Compare $\tilde{\beta}$ from the OLS regresion

$$Y^* = X^*\tilde{\beta} + \tilde{e}$$

with $\hat{\beta}$ from the OLS regression

$$Y = D_1\hat{\alpha}_1 + D_2\hat{\alpha}_2 + X\hat{\beta} + \hat{e}.$$

12. The data file *cps85.dat* contains a random sample of 528 individuals from the 1985 Current Population Survey by the U.S. Census Bureau. The file contains observations on nine variables, listed in the file *cps85.pdf*.

V1 = education (in years)
V2 = region of residence (coded 1 if South, 0 otherwise)
V3 = (coded 1 if nonwhite and non-Hispanic, 0 otherwise)
V4 = (coded 1 if Hispanic, 0 otherwise)
V5 = gender (coded 1 if female, 0 otherwise)
V6 = marital status (coded 1 if married, 0 otherwise)
V7 = potential labor market experience (in years)
V8 = union status (coded 1 if in union job, 0 otherwise)
V9 = hourly wage (in dollars)

Estimate a regression of wage y_i on education x_{1i}, experience x_{2i}, and experienced-squared $x_{3i} = x_{2i}^2$ (and a constant). Report the OLS estimates.

Let $\hat{e}_i$ be the OLS residual and $\hat{y}_i$ the predicted value from the regression. Numerically calculate the following:

(a) $\sum_{i=1}^n \hat{e}_i$
(b) $\sum_{i=1}^n x_{1i}\hat{e}_i$
(c) $\sum_{i=1}^n x_{2i}\hat{e}_i$
(d) $\sum_{i=1}^n x_{1i}^2\hat{e}_i$
(e) $\sum_{i=1}^n x_{2i}^2\hat{e}_i$

(f) $\sum_{i=1}^{n} \hat{y}_i \hat{e}_i$

(g) $\sum_{i=1}^{n} \hat{e}_i^2$

(h) R^2

Are the calculations (i)-(vi) consistent with the theoretical properties of OLS? Explain.

13. Use the data from the previous problem, restimate the slope on education using the residual regression approach. Regress y_i on $(1, x_{2i}, x_{2i}^2)$, regress x_{1i} on $(1, x_{2i}, x_{2i}^2)$, and regress the residuals on the residuals. Report the estimate from this regression. Does it equal the value from the first OLS regression? Explain.

    In the second-stage residual regression, (the regression of the residuals on the residuals), calculate the equation R^2 and sum of squared errors. Do they equal the values from the initial OLS regression? Explain.

Chapter 5

Asymptotic Theory

This chapter reviews the essential components of asymptotic theory.

5.1 Inequalities

Asymptotic theory is based on a set of approximations. These approximations are bounded through the use of mathematical inequalities. We list here some of the most critical definitions and inequalities.

The **Euclidean norm** of an $m \times 1$ vector a is

$$|a| = \left(a'a\right)^{1/2} = \left(\sum_{i=1}^{m} a_i^2\right)^{1/2}.$$

If A is a $m \times n$ matrix, then its Euclidean norm is

$$|A| = \operatorname{tr}\left(A'A\right)^{1/2} = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{1/2}.$$

The following are an important set of inequalities which are used in asymptotic distribution theory.

Triangle inequality

$$|X + Y| \leq |X| + |Y|.$$

Jensen's Inequality. If $g(\cdot) : R \to R$ is convex, then

$$g(E(X)) \leq E(g(X)). \tag{5.1}$$

Cauchy-Schwarz Inequality.

$$E|XY| \leq \left(E|X|^2\right)^{1/2}\left(E|Y|^2\right)^{1/2} \tag{5.2}$$

Holder's Inequality. If $p > 1$ and $q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, then

$$E|XY| \leq \left(E|X|^p\right)^{1/p}\left(E|Y|^q\right)^{1/q}. \tag{5.3}$$

Markov's Inequality. For any strictly increasing function $g(X) \geq 0$,

$$P(g(X) > \alpha) \leq \alpha^{-1} E g(X). \tag{5.4}$$

Proof of Jensen's Inequality. Let $a + bx$ be the tangent line to $g(x)$ at $x = EX$. Since $g(x)$ is convex, tangent lines lie below it. So for all x, $g(x) \geq a + bx$ yet $g(EX) = a + bEX$ since the curve is tangent at EX. Applying expectations, $Eg(X) \geq a + bEX = g(EX)$, as stated. ■

Proof of Holder's Inequality. Let $U = |X|^p / E|X|^p$ and $V = |Y|^q / E|Y|^q$. Note $EU = EV = 1$. Since $\frac{1}{p} + \frac{1}{q} = 1$ an application of Jensen's inequality shows that

$$U^{1/p} V^{1/q} = \exp\left[\frac{1}{p}\ln U + \frac{1}{q}\ln V\right] \leq \frac{1}{p}\exp(\ln U) + \frac{1}{p}\exp(\ln V) = \frac{U}{p} + \frac{V}{q}.$$

Then

$$\frac{E|XY|}{(E|X|^p)^{1/p}(E|Y|^q)^{1/q}} = E\left(U^{1/p}V^{1/q}\right) \leq E\left(\frac{U}{p} + \frac{V}{q}\right) = \frac{1}{p} + \frac{1}{q} = 1,$$

which is (5.3). ■

Proof of Markov's Inequality. Set $Y = g(X)$ and let f denote the density function of Y. Then

$$\begin{aligned} P(Y > \alpha) &= \alpha^{-1} \int_{\alpha}^{\infty} \alpha f(y) dy \\ &\leq \alpha^{-1} \int_{\alpha}^{\infty} y f(y) dy \\ &\leq \alpha^{-1} \int_{-\infty}^{\infty} y f(y) dy = \alpha^{-1} E(Y) \end{aligned}$$

the second-to-last inequality using the region of integration $\{y > \alpha\}$. ■

5.2 Weak Law of Large Numbers

Let $Z_n \in R^k$ be a random vector. We say that Z_n **converges in probability** to Z as $n \to \infty$, denoted $Z_n \to_p Z$ as $n \to \infty$, if for all $\delta > 0$,

$$\lim_{n\to\infty} P(|Z_n - Z| > \delta) = 0.$$

This is a probabilistic way of generalizing the mathematical definition of a limit. The WLLN shows that sample averages converge in probability to the population average.

Theorem 5.2.1 ***Weak Law of Large Numbers (WLLN)***. *If $X_i \in R^k$ is iid and $E|X_i| < \infty$, then as $n \to \infty$*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to_p E(X).$$

Proof: Without loss of generality, we can set $E(X) = 0$ (by recentering X_i on its expectation). We need to show that for all $\delta > 0$ and $\eta > 0$ there is some $N < \infty$ so that for all $n \geq N$, $P\left(\left|\overline{X}_n\right| > \delta\right) \leq \eta$. Fix δ and η. Set $\varepsilon = \delta\eta/3$. Pick $C < \infty$ large enough so that

$$E\left(|X| 1\left(|X| > C\right)\right) \leq \varepsilon \tag{5.5}$$

(where $1(\cdot)$ is the indicator function) which is possible since $E|X| < \infty$. Define the random vectors

$$\begin{aligned} W_i &= X_i 1\left(|X_i| \leq C\right) - E\left(X_i 1\left(|X_i| \leq C\right)\right) \\ Z_i &= X_i 1\left(|X_i| > C\right) - E\left(X_i 1\left(|X_i| > C\right)\right). \end{aligned}$$

By the triangle inequality, Jensen's inequality (5.1) and (5.5),

$$\begin{aligned} E\left|\overline{Z}_n\right| &\leq E|Z_i| \\ &\leq E|X_i| 1\left(|X_i| > C\right) + \left|E\left(X_i 1\left(|X_i| > C\right)\right)\right| \\ &\leq 2E|X_i| 1\left(|X_i| > C\right) \\ &\leq 2\varepsilon. \end{aligned} \tag{5.6}$$

By Jensen's inequality (5.1), the fact that the W_i are iid and mean zero, and the bound $|W_i| \leq 2C$,

$$\begin{aligned} \left(E\left|\overline{W}_n\right|\right)^2 &\leq E\overline{W}_n^2 \\ &= \frac{EW_i^2}{n} \\ &\leq \frac{4C^2}{n} \\ &\leq \varepsilon^2 \end{aligned} \tag{5.7}$$

the final inequality holding for $n \geq 4C^2/\varepsilon^2 = 36C^2/\delta^2\eta^2$.

Finally, by Markov's inequality (5.4), the fact that $\overline{X}_n = \overline{W}_n + \overline{Z}_n$, the triangle inequality, (5.6) and (5.7),

$$P\left(\left|\overline{X}_n\right| > \delta\right) \leq \frac{E\left|\overline{X}_n\right|}{\delta} \leq \frac{E\left|\overline{W}_n\right| + E\left|\overline{Z}_n\right|}{\delta} \leq \frac{3\varepsilon}{\delta} = \eta,$$

the equality by the definition of ε. We have shown that for any $\delta > 0$ and $\eta > 0$ then for all $n \geq 36C^2/\delta^2\eta^2$, $P\left(\left|\overline{X}_n\right| > \delta\right) \leq \eta$, as needed. ■

5.3 Convergence in Distribution

Let Z_n be a random variable with distribution $F_n(x) = P\left(Z_n \leq x\right)$. We say that Z_n **converges in distribution** to Z as $n \to \infty$, denoted $Z_n \to_d Z$, where Z has distribution $F(x) = P\left(Z \leq x\right)$, if for all x at which $F(x)$ is continuous, $F_n(x) \to F(x)$ as $n \to \infty$.

Theorem 5.3.1 ***Central Limit Theorem (CLT).*** *If $X_i \in R^k$ is iid and $E|X_i|^2 < \infty$, then as $n \to \infty$*

$$\sqrt{n}\left(\overline{X}_n - \mu\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(X_i - \mu\right) \to_d N\left(0, V\right).$$

where $\mu = EX$ and $V = E\left(X - \mu\right)\left(X - \mu\right)'$.

Proof: Without loss of generality, it is sufficient to consider the case $\mu = 0$ and $V = I_k$. For $\lambda \in R^k$, let $C(\lambda) = E\exp\left(i\lambda' X\right)$ denote the characteristic function of X and set $c(\lambda) = \ln C(\lambda)$. Then observe

$$\begin{aligned}\frac{\partial}{\partial\lambda}C(\lambda) &= iE\left(X\exp\left(i\lambda' X\right)\right)\\ \frac{\partial^2}{\partial\lambda\partial\lambda'}C(\lambda) &= i^2E\left(XX'\exp\left(i\lambda' X\right)\right)\end{aligned}$$

so when evaluated at $\lambda = 0$

$$\begin{aligned}C(0) &= 1\\ \frac{\partial}{\partial\lambda}C(0) &= iE\left(X\right) = 0\\ \frac{\partial^2}{\partial\lambda\partial\lambda'}C(0) &= -E\left(XX'\right) = -I_k.\end{aligned}$$

Furthermore,

$$\begin{aligned}c_\lambda(\lambda) &= \frac{\partial}{\partial\lambda}c(\lambda) = C(\lambda)^{-1}\frac{\partial}{\partial\lambda}C(\lambda)\\ c_{\lambda\lambda}(\lambda) &= \frac{\partial^2}{\partial\lambda\partial\lambda'}c(\lambda) = C(\lambda)^{-1}\frac{\partial^2}{\partial\lambda\partial\lambda'}C(\lambda) - C(\lambda)^{-2}\frac{\partial}{\partial\lambda}C(\lambda)\frac{\partial}{\partial\lambda'}C(\lambda)\end{aligned}$$

so when evaluated at $\lambda = 0$

$$\begin{aligned}c(0) &= 0\\ c_\lambda(0) &= 0\\ c_{\lambda\lambda}(0) &= -I_k.\end{aligned}$$

By a second-order Taylor series expansion of $c(\lambda)$ about $\lambda = 0$,

$$c(\lambda) = c(0) + c_\lambda(0)'\lambda + \frac{1}{2}\lambda' c_{\lambda\lambda}(\lambda^*)\lambda = \frac{1}{2}\lambda' c_{\lambda\lambda}(\lambda^*)\lambda \tag{5.8}$$

where λ^* lies on the line segment joining 0 and λ.

We now compute $C_n(\lambda) = E\exp\left(i\lambda'\sqrt{n}\overline{X}_n\right)$ the characteristic function of $\sqrt{n}\overline{X}_n$. By the properties of the exponential function, the independence of the X_i, the definition of $c(\lambda)$ and (5.8)

$$\begin{aligned}\ln C_n(\lambda) &= \log E\exp\left(i\frac{1}{\sqrt{n}}\sum_{j=1}^{n\lambda\prime}X_j\right)\\ &= \log E\prod_{j=1}^{n}\exp\left(i\frac{1}{\sqrt{n}}\lambda' X_j\right)\\ &= \log\prod_{i=1}^{n}E\exp\left(i\frac{1}{\sqrt{n}}\lambda' X_j\right)\\ &= nc\left(\frac{\lambda}{\sqrt{n}}\right)\\ &= \frac{1}{2}\lambda' c_{\lambda\lambda}(\lambda_n)\lambda\end{aligned}$$

where $\lambda_n \to 0$ lies on the line segment joining 0 and $\lambda/\sqrt{n}$. Since $c_{\lambda\lambda}(\lambda_n) \to c_{\lambda\lambda}(0) = -I_k$, we see that as $n \to \infty$,

$$C_n(\lambda) \to \exp\left(-\frac{1}{2}\lambda'\lambda\right)$$

the characteristic function of the $N(0, I_k)$ distribution. This is sufficient to establish the theorem. ■

5.4 Asymptotic Transformations

Theorem 5.4.1 ***Continuous Mapping Theorem 1*** *(CMT). If $Z_n \to_p c$ as $n \to \infty$ and $g(\cdot)$ is continuous at c, then $g(Z_n) \to_p g(c)$ as $n \to \infty$.*

Proof: Since g is continuous at c, for all $\varepsilon > 0$ we can find a $\delta > 0$ such that if $|Z_n - c| < \delta$ then $|g(Z_n) - g(c)| \leq \varepsilon$. Recall that $A \subset B$ implies $P(A) \leq P(B)$. Thus $P(|g(Z_n) - g(c)| \leq \varepsilon) \geq P(|Z_n - c| < \delta) \to 1$ as $n \to \infty$ by the assumption that $Z_n \to_p c$. Hence $g(Z_n) \to_p g(c)$ as $n \to \infty$.

Theorem 5.4.2 ***Continuous Mapping Theorem 2.*** *If $Z_n \to_d Z$ as $n \to \infty$ and $g(\cdot)$ is continuous, then $g(Z_n) \to_d g(Z)$ as $n \to \infty$.*

Theorem 5.4.3 ***Delta Method:*** *If $\sqrt{n}(\theta_n - \theta_0) \to_d N(0, \Sigma)$, where θ is $m \times 1$ and Σ is $m \times m$, and $g(\theta) : R^m \to R^k$, $k \leq m$, then*

$$\sqrt{n}(g(\theta_n) - g(\theta_0)) \to_d N\left(0, g_\theta \Sigma g_\theta'\right)$$

where $g_\theta(\theta) = \frac{\partial}{\partial \theta'} g(\theta)$ and $g_\theta = g_\theta(\theta_0)$.

Proof: By a vector Taylor series expansion, for each element of g,

$$g_j(\theta_n) = g_j(\theta_0) + g_{j\theta}(\theta_{jn}^*)(\theta_n - \theta_0)$$

where θ_{nj} lies on the line segment between θ_n and θ_0 and therefore converges in probability to θ_0. It follows that $a_{jn} = g_{j\theta}(\theta_{jn}^*) - g_{j\theta} \to_p 0$. Stacking across elements of g, we find

$$\sqrt{n}(g(\theta_n) - g(\theta_0)) = (g_\theta + a_n)\sqrt{n}(\theta_n - \theta_0) \to_d g_\theta N(0, \Sigma) = N\left(0, g_\theta \Sigma g_\theta'\right).$$

Chapter 6

Inference

6.1 Sampling Distribution

The least-squares estimator is a random vector, since it is a function of the random data, and therefore has a sampling distribution. In general, its distribution is a complicated function of the joint distribution of (y_i, x_i) and the sample size n.

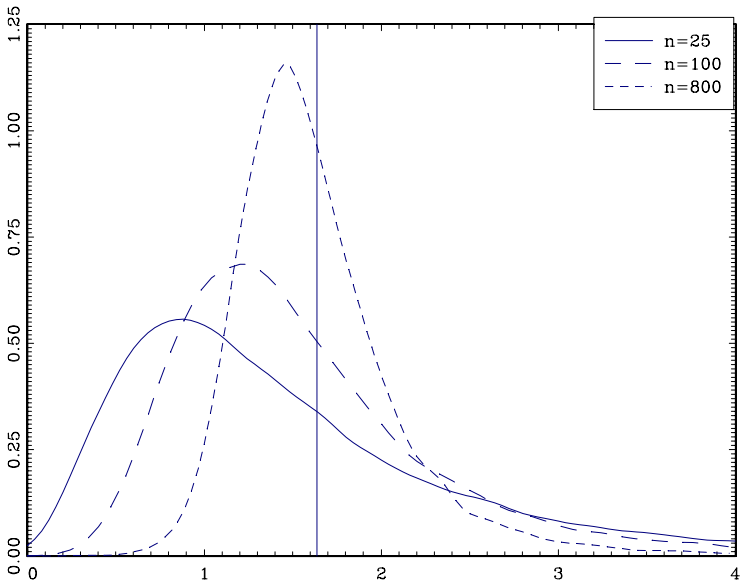

Figure 6.1: Sampling Density of $\hat{\beta}_2$

To illustrate the possibilities in one example, let y_i and x_i be drawn from the joint density

$$f(x,y) = \frac{1}{2\pi xy} \exp\left(-\frac{1}{2}\left(\ln y - \ln x\right)^2\right) \exp\left(-\frac{1}{2}\left(\ln x\right)^2\right)$$

and let $\hat{\beta}_2$ be the slope coefficient estimate computed on observations from this joint density. Using simulation methods, the density function of $\hat{\beta}_2$ was computed and plotted in Figure 6.1 for sample sizes of $n = 25$, $n = 100$ and $n = 800$. The verticle line marks the true value of the projection coefficient.

From the figure we can see that the density functions are dispersed and highly non-normal. As the sample size increases the density becomes more concentrated about the population coefficient. To characterize the sampling distribution more fully, we will use the methods of asymptotic approximation.

6.2 Consistency

As discussed in Section 6.1, the OLS estimator $\hat{\beta}$ is has a statistical distribution which is unknown. Asymptotic (large sample) methods approximate sampling distributions based on the limiting experiment that the sample size n tends to infinity. A preliminary step in this approach is the demonstration that estimators are consistent – that they converge in probability to the true parameters as the sample size gets large. This is illustrated in Figure 6.1 by the fact that the sampling densities become more concentrated as n gets larger.

Theorem 6.2.1 *Under Assumption 3.5.1,* $\hat{\beta} \to_p \beta$ *as* $n \to \infty$.

Proof. Equation (4.8) implies that

$$\hat{\beta} - \beta = \left(\sum_{i=1}^{n} x_i x_i' \right)^{-1} \sum_{i=1}^{n} x_i e_i. \tag{6.1}$$

We now deduce the consistency of $\hat{\beta}$. First, Assumption 3.5.1 and the WLLN (Theorem 5.2.1) imply that

$$\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \to_p E\left(x_i x_i'\right) = Q \tag{6.2}$$

and

$$\frac{1}{n} \sum_{i=1}^{n} x_i e_i \to_p E\left(x_i e_i\right) = 0. \tag{6.3}$$

From (6.1), (6.2), (6.3), and the continuous mapping theorem (Theorem 5.4.1), we can conclude that $\hat{\beta} \to_p \beta$. For a complete argument, using (6.1), we can write

$$\begin{aligned} \hat{\beta} - \beta &= \left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i e_i \right) \\ &= g\left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i', \frac{1}{n} \sum_{i=1}^{n} x_i e_i \right) \end{aligned}$$

where $g(A, b) = A^{-1} b$ is a continuous function of A and b at all values of the arguments such that A^{-1} exist. Assumption 3.5.1.4 implies that Q^{-1} exists and thus $g(\cdot, \cdot)$ is continuous at $(Q, 0)$. Hence by the continuous mapping theorem (Theorem 5.4.1),

$$\hat{\beta} - \beta = g\left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i', \frac{1}{n} \sum_{i=1}^{n} x_i e_i \right) \to_p g\left(Q, 0\right) = Q^{-1} 0 = 0$$

which implies $\hat{\beta} \to_p \beta$ as stated. ■

We can similarly show that the estimators $\hat{\sigma}^2$ and s^2 are consistent for σ^2.

Theorem 6.2.2 *Under Assumption 3.5.1, $\hat{\sigma}^2 \to_p \sigma^2$ and $s^2 \to_p \sigma^2$ as $n \to \infty$.*

Proof. Note that

$$\begin{aligned} \hat{e}_i &= y_i - x_i'\hat{\beta} \\ &= e_i + x_i'\beta - x_i'\hat{\beta} \\ &= e_i - x_i'\left(\hat{\beta} - \beta\right). \end{aligned}$$

Thus

$$\hat{e}_i^2 = e_i^2 - 2e_i x_i'\left(\hat{\beta} - \beta\right) + \left(\hat{\beta} - \beta\right)' x_i x_i'\left(\hat{\beta} - \beta\right) \tag{6.4}$$

and

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n \hat{e}_i^2 \\ &= \frac{1}{n}\sum_{i=1}^n e_i^2 - 2\left(\frac{1}{n}\sum_{i=1}^n e_i x_i'\right)\left(\hat{\beta} - \beta\right) + \left(\hat{\beta} - \beta\right)'\left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)\left(\hat{\beta} - \beta\right) \\ &\to_p \sigma^2 \end{aligned}$$

the last line using the WLLN, (6.2), (6.3) and Theorem (6.2.1). Thus $\hat{\sigma}^2$ is consistent for σ^2.

Finally, since $n/(n-k) \to 1$ as $n \to \infty$, it follows that

$$s^2 = \frac{n}{n-k}\hat{\sigma}^2 \to_p \sigma^2.$$

■

6.3 Asymptotic Normality

We now establish the asymptotic distribution of $\hat{\beta}$ after normalization. We need a strengthening of the moment conditions.

Assumption 6.3.1 *In addition to Assumption 3.5.1, $Ee_i^4 < \infty$ and $E\left|x_i\right|^4 < \infty$.*

Now define

$$\Omega = E\left(x_i x_i' e_i^2\right).$$

Assumption 6.3.1 guarantees that the elements of Ω are finite. To see this, by the Cauchy-Schwarz inequality (5.2),

$$E\left|x_i x_i' e_i^2\right| \leq \left(E\left|x_i x_i'\right|^2\right)^{1/2}\left(E\left|e_i^4\right|\right)^{1/2} = \left(E\left|x_i\right|^4\right)^{1/2}\left(E\left|e_i^4\right|\right)^{1/2} < \infty. \tag{6.5}$$

Thus $x_i e_i$ is iid with mean zero and has covariance matrix Ω. By the central limit theorem (Theorem 5.3.1),

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n x_i e_i \to_d N\left(0, \Omega\right). \tag{6.6}$$

Then using (6.1), (6.2), and (6.6),

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}-\beta\right) &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i e_i\right) \\ &\to_d Q^{-1}N\left(0,\Omega\right) \\ &= N\left(0, Q^{-1}\Omega Q^{-1}\right).\end{aligned}$$

Theorem 6.3.1 *Under Assumption 6.3.1, as* $n \to \infty$

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \to_d N\left(0, V\right)$$

where $V = Q^{-1}\Omega Q^{-1}$.

As V is the variance of the asymptotic distribution of $\sqrt{n}\left(\hat{\beta}-\beta\right)$, V is often referred to as the **asymptotic covariance matrix** of $\hat{\beta}$. The expression $V = Q^{-1}\Omega Q^{-1}$ is called a **sandwich** form.

Theorem 6.3.1 states that the sampling distribution of the least-squares estimator, after rescaling, is approximately normal when the sample size n is sufficiently large. This holds true for all joint distibutions of (y_i, x_i) which satisfy the conditions of Assumption 6.3.1. However, for any fixed n the sampling distribution of $\hat{\beta}$ can be arbitrarily far from the normal distribution. In Figure 6.1 we have already seen a simple example where the least-squares estimate is quite asymmetric and non-normal even for reasonably large sample sizes.

There is a special case where Ω and V simplify. We say that e_i is a **Homoskedastic Projection Error** when

$$Cov(x_i x_i', e_i^2) = 0. \tag{6.7}$$

Condition (6.7) holds, for example, when x_i and e_i are independent, but this is not a necessary condition. Under (6.7) the asymptotic variance formulas simplify as

$$\Omega = E\left(x_i x_i'\right) E\left(e_i^2\right) = Q\sigma^2 \tag{6.8}$$

$$V = Q^{-1}\Omega Q^{-1} = Q^{-1}\sigma^2 \equiv V^0 \tag{6.9}$$

In (6.9) we define $V^0 = Q^{-1}\sigma^2$ whether (6.7) is true or false. When (6.7) is true then $V = V^0$, otherwise $V \neq V^0$. We call V^0 the **homoskedastic covariance matrix**.

The asymptotic distribution of Theorem 6.3.1 is commonly used to approximate the finite sample distribution of $\sqrt{n}\left(\hat{\beta}-\beta\right)$. The approximation may be poor when n is small. How large should n be in order for the approximation to be useful? Unfortunately, there is no simple answer to this reasonable question. The trouble is that no matter how large is the sample size, the normal approximation is arbitrarily poor for some data distribution satisfying the assumptions. We illustrate this problem using a simulation. Let $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where x_i is $N(0,1)$, and ε_i is independent of x_i with the Double Pareto density $f(\varepsilon) = \frac{\alpha}{2}|\varepsilon|^{-\alpha-1}$, $|\varepsilon| \geq 1$. If $\alpha > 2$ the error ε_i has zero mean and variance $\alpha/(\alpha-2)$. As α approaches 2, however, its variance diverges to infinity. In this context the normalized least-squares slope estimator $\sqrt{n\frac{\alpha-2}{\alpha}}\left(\hat{\beta}_2-\beta_2\right)$ has the $N(0,1)$ asymptotic distibution. In Figure 6.2 we display the finite sample densities of the normalized estimator $\sqrt{n\frac{\alpha-2}{\alpha}}\left(\hat{\beta}_2-\beta_2\right)$, setting $n = 100$ and varying the parameter α. For

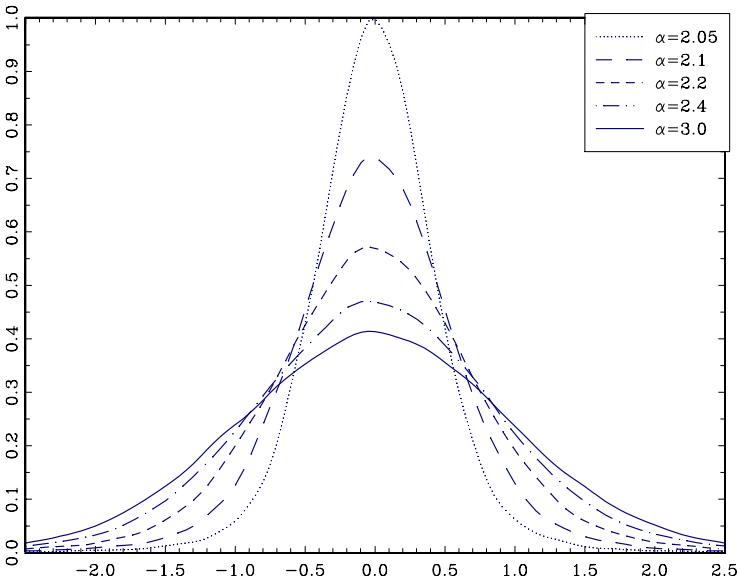

Figure 6.2: Density of Normalized OLS estimator

$\alpha = 3.0$ the density is very close to the $N(0,1)$ density. As α diminishes the density changes significantly, concentrating most of the probability mass around zero.

Another example is shown in Figure 6.3. Here the model is $y_i = \beta_1 + \varepsilon_i$ where

$$\varepsilon_i = \frac{u_i^k - Eu_i^k}{\left(Eu_i^{2k} - \left(Eu_i^k\right)^2\right)^{1/2}}$$

and $u_i \sim N(0,1)$. We show the sampling distribution of $\sqrt{n}\left(\hat{\beta}_1 - \beta_1\right)$ setting $n = 100$, for $k = 1$, 4, 6 and 8. As k increases, the sampling distribution becomes highly skewed and non-normal. The lesson from Figures 6.2 and 6.3 is that the $N(0,1)$ asymptotic approximation is never guaranteed to be accurate.

6.4 Covariance Matrix Estimation

Let

$$\hat{Q} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i'$$

be the method of moments estimator for Q. The homoskedastic covariance matrix $V^0 = Q^{-1}\sigma^2$ is typically estimated by

$$\hat{V}^0 = \hat{Q}^{-1}s^2. \tag{6.10}$$

Since $\hat{Q} \to_p Q$ and $s^2 \to_p \sigma^2$ (see (6.2) and Theorem 6.2.1) it is clear that $\hat{V}^0 \to_p V^0$. The estimator $\hat{\sigma}^2$ may also be substituted for s^2 in (6.10) without changing this result.

To estimate $V = Q^{-1}\Omega Q^{-1}$, we need an estimate of $\Omega = E\left(x_i x_i' e_i^2\right)$. The MME estimator is

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i' \hat{e}_i^2 \tag{6.11}$$

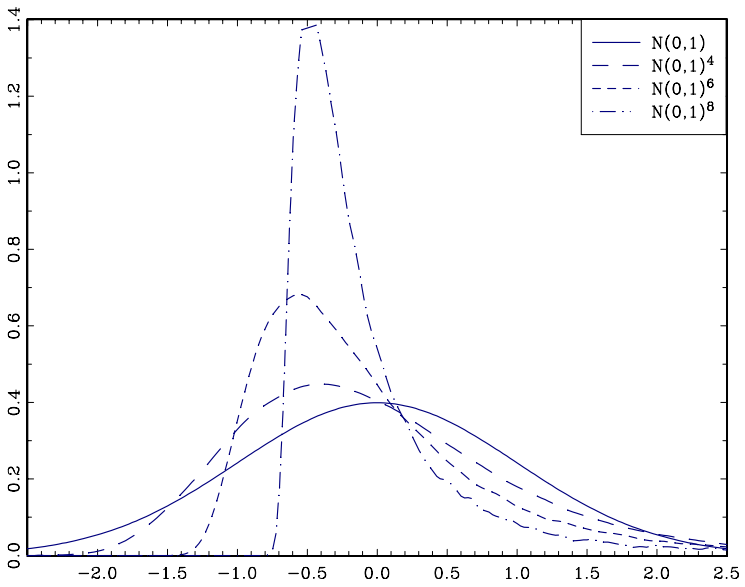

Figure 6.3: Sampling distribution

where $\hat{e}_i$ are the OLS residuals. The estimator of V is then

$$\hat{V} = \hat{Q}^{-1}\hat{\Omega}\hat{Q}^{-1}$$

This estimator was introduced to the econometrics literature by White (1980).

The estimator $\hat{V}^0$ was the dominate covariance estimator used before 1980, and was still the standard choice for much empirical work done in the early 1980s. The methods switched during the late 1980s and early 1990s, so that by the late 1990s White estimate $\hat{V}$ emerged as the standard covariance matrix estimator. When reading and reporting applied work, it is important to pay attention to the distinction between $\hat{V}^0$ and $\hat{V}$, as it is not always clear which has been computed. When $\hat{V}$ is used rather than the traditional choice $\hat{V}^0$, many authors will state that their "standard errors have been corrected for heteroskedasticity", or that they use a "heteroskedasticity-robust covariance matrix estimator", or that they use the "White formula", the "Eicker-White formula", the "Huber formula", the "Huber-White formula" or the "GMM covariance matrix". In most cases, these all mean the same thing.

The variance estimator $\hat{V}$ is an estimate of the variance of the asymptotic distribution of $\hat{\beta}$. A more easily interpretable measure of spread is its square root – the standard deviation. This motivates the definition of a standard error.

Definition 6.4.1 *A **standard error** $s(\hat{\beta})$ for an estimator $\hat{\beta}$ is an estimate of the standard deviation of the distribution of $\hat{\beta}$.*

When β is scalar, and $\hat{V}$ is an estimator of the variance of $\sqrt{n}\left(\hat{\beta} - \beta\right)$, we set $s(\hat{\beta}) = n^{-1/2}\sqrt{\hat{V}}$. When β is a vector, we focus on individual elements of β one-at-a-time, vis., β_j, $j = 0, 1, ..., k$. Thus

$$s(\hat{\beta}_j) = n^{-1/2}\sqrt{\hat{V}_{jj}}.$$

Generically, standard errors are not unique, as there may be more than one estimator of the variance of the estimator. It is therefore important to understand what formula and method is

used by an author when studying their work. It is also important to understand that a particular standard error may be relevant under one set of model assumptions, but not under another set of assumptions, just as any other estimator.

From a computational standpoint, the standard method to calculate the standard errors is to first calculate $n^{-1}\hat{V}$, then take the diagonal elements, and then the square roots.

To illustrate, we return to the log wage regression of Section 4.1. We calculate that $s^2 = 0.20$ and

$$\hat{\Omega} = \begin{pmatrix} 0.199 & 2.80 \\ 2.80 & 40.6 \end{pmatrix}.$$

Therefore the two covariance matrix estimates are

$$\hat{V}^0 = \begin{pmatrix} 1 & 14.14 \\ 14.14 & 205.83 \end{pmatrix}^{-1} 0.20 = \begin{pmatrix} 6.98 & -0.480 \\ -0.480 & .039 \end{pmatrix}$$

and

$$\hat{V} = \begin{pmatrix} 1 & 14.14 \\ 14.14 & 205.83 \end{pmatrix}^{-1} \begin{pmatrix} .199 & 2.80 \\ 2.80 & 40.6 \end{pmatrix} \begin{pmatrix} 1 & 14.14 \\ 14.14 & 205.83 \end{pmatrix}^{-1} = \begin{pmatrix} 7.20 & -0.493 \\ -0.493 & 0.035 \end{pmatrix}.$$

In this case the two estimates are quite similar. The standard errors for $\hat{\beta}_0$ are $\sqrt{7.2/988} = .085$ and that for $\hat{\beta}_1$ is $\sqrt{.35/988} = .020$. We can write the estimated equation with standards errors using the format

$$\begin{array}{rccc} \widehat{\log(Wage_i)} = & 1.30 & + & 0.117 \quad Education_i. \\ & (.085) & & (.020) \end{array}$$

6.5 Consistency of the White Covariance Matrix Estimate

We now show $\hat{\Omega} \to_p \Omega$, from which it follows that $\hat{V} \to_p V$ as $n \to \infty$. Using (6.4)

$$\begin{aligned} \hat{\Omega} &= \frac{1}{n}\sum_{i=1}^{n} x_i x_i' \hat{e}_i^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i x_i' e_i^2 - \frac{2}{n}\sum_{i=1}^{n} x_i x_i' \left(\hat{\beta} - \beta\right)' x_i e_i + \frac{1}{n}\sum_{i=1}^{n} x_i x_i' \left(\hat{\beta} - \beta\right)' x_i x_i' \left(\hat{\beta} - \beta\right). \end{aligned} \quad (6.12)$$

We now examine each sum on the right-hand-side of (6.12) in turn. First, (6.5) and the WLLN (Theorem 5.2.1) show that

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' e_i^2 \to_p E\left(x_i x_i' e_i^2\right) = \Omega.$$

Second, by Holder's inequality (5.3)

$$E\left(|x_i|^3 |e_i|\right) \le \left(E|x_i|^4\right)^{3/4} \left(E\left|e_i^4\right|\right)^{1/4} < \infty,$$

so by the WLLN

$$\frac{1}{n}\sum_{i=1}^{n} |x_i|^3 |e_i| \to_p E\left(|x_i|^3 |e_i|\right),$$

and thus since $\left|\hat{\beta}-\beta\right| \to_p 0$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \left(\hat{\beta}-\beta\right)' x_i e_i\right| \leq \left|\hat{\beta}-\beta\right| \left(\frac{1}{n}\sum_{i=1}^{n} |x_i|^3 |e_i|\right) \to_p 0.$$

Third, by the WLLN

$$\frac{1}{n}\sum_{i=1}^{n} |x_i|^4 \to_p E|x_i|^4,$$

so

$$\left|\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \left(\hat{\beta}-\beta\right)' x_i x_i \left(\hat{\beta}-\beta\right)\right| \leq \left|\hat{\beta}-\beta\right|^2 \frac{1}{n}\sum_{i=1}^{n} |x_i|^4 \to_p 0.$$

Together, these establish consistency.

Theorem 6.5.1 *As $n \to \infty$, $\hat{\Omega} \to_p \Omega$ and $\hat{V} \to_p V$.*

6.6 Alternative Covariance Matrix Estimators

MacKinnon and White (1985) suggested a small-sample corrected version of $\hat{V}$ based on the jackknife principle. Recall from Section 4.12 the definition of $\hat{\beta}_{(-i)}$ as the least-squares estimator with the i'th observation deleted. From equation (3.13) of Efron (1982), the jackknife estimator of the variance matrix for $\hat{\beta}$ is

$$\hat{V}^* = (n-1)\sum_{i=1}^{n} \left(\hat{\beta}_{(-i)} - \bar{\beta}\right)\left(\hat{\beta}_{(-i)} - \bar{\beta}\right)' \tag{6.13}$$

where

$$\bar{\beta} = \frac{1}{n}\sum_{i=1}^{n} \hat{\beta}_{(-i)}.$$

Using formula (4.25), you can show that

$$\hat{V}^* = \frac{n-1}{n}\hat{Q}^{-1}\hat{\Omega}^*\hat{Q}^{-1} \tag{6.14}$$

where

$$\hat{\Omega}^* = \frac{1}{n}\sum_{i=1}^{n} (1-h_i)^{-2} x_i x_i' \hat{e}_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} (1-h_i)^{-1} x_i \hat{e}_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} (1-h_i)^{-1} x_i \hat{e}_i\right)'$$

and $h_i = x_i' (X'X)^{-1} x_i$. MacKinnon and White (1985) present numerical (simulation) evidence that $\hat{V}^*$ works better than $\hat{V}$ as an estimator of V. They also suggest that the scaling factor $(n-1)/n$ in (**??**) can be omitted.

Andrews (1991) suggested an similar estimator based on cross-validation, which. is defined by replacing the OLS residual $\hat{e}_i$ in (6.11) with the leave-one-out estimator $\hat{e}_{i,-i} = (1-h_i)^{-1}\hat{e}_i$ presented in (4.26). Using this substitution, Andrews' proposed estimator is

$$\hat{V}^{**} = \hat{Q}^{-1}\hat{\Omega}^{**}\hat{Q}^{-1}$$

where

$$\hat{\Omega}^{**} = \frac{1}{n}\sum_{i=1}^{n}(1-h_i)^{-2}\,x_i x_i' \hat{e}_i^2.$$

It is similar to the MacKinnon-White estimator $\hat{V}^*$, but omits the mean correction. Andrews (1991) argues that simulation evidence indicates that $\hat{V}^{**}$ is an improvement on $\hat{V}^*$.

6.7 Functions of Parameters

Sometimes we are interested in some lower-dimensional function of the parameter vector $\beta = (\beta_1, ..., \beta_{k+1})$. For example, we may be interested in a single coefficient β_j or a ratio β_j/β_l. In these cases we can write the parameter of interest as a function of β. Let $h : R^k \to R^q$ denote this function and let

$$\theta = h(\beta)$$

denote the parameter of interest. The estimate of θ is

$$\hat{\theta} = h(\hat{\beta}).$$

What is an appropriate standard error for $\hat{\theta}$? Assume that $h(\beta)$ is differentiable at the true value of β. By a first-order Taylor series approximation:

$$h(\hat{\beta}) \simeq h(\beta) + H_\beta' \left(\hat{\beta} - \beta\right).$$

where

$$H_\beta = \frac{\partial}{\partial \beta} h(\beta) \qquad (k+1) \times q.$$

Thus

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta} - \theta\right) &= \sqrt{n}\left(h(\hat{\beta}) - h(\beta)\right) \\
&\simeq H_\beta' \sqrt{n}\left(\hat{\beta} - \beta\right) \\
&\to_d H_\beta' N(0, V) \\
&= N(0, V_\theta). \qquad (6.15)
\end{aligned}$$

where

$$V_\theta = H_\beta' V H_\beta.$$

If $\hat{V}$ is the estimated covariance matrix for $\hat{\beta}$, then the natural estimate for the variance of $\hat{\theta}$ is

$$\hat{V}_\theta = \hat{H}_\beta' \hat{V} \hat{H}_\beta$$

where

$$\hat{H}_\beta = \frac{\partial}{\partial \beta} h(\hat{\beta}).$$

In many cases, the function $h(\beta)$ is linear:

$$h(\beta) = R'\beta$$

for some $k \times q$ matrix R. In this case, $H_\beta = R$ and $\hat{H}_\beta = R$, so $\hat{V}_\theta = R'\hat{V}R$.

For example, if R is a "selector matrix"

$$R = \begin{pmatrix} I \\ 0 \end{pmatrix}$$

so that if $\beta = (\beta_1, \beta_2)$, then $\theta = R'\beta = \beta_1$ and

$$\hat{V}_\theta = \begin{pmatrix} I & 0 \end{pmatrix} \hat{V} \begin{pmatrix} I \\ 0 \end{pmatrix} = \hat{V}_{11},$$

the upper-left block of $\hat{V}$.

When $q = 1$ (so $h(\beta)$ is real-valued), the standard error for $\hat{\theta}$ is the square root of $n^{-1}\hat{V}_\theta$, that is, $s(\hat{\theta}) = n^{-1/2}\sqrt{\hat{H}'_\beta \hat{V} \hat{H}_\beta}$.

6.8 t tests

Let $\theta = h(\beta) : R^k \to R$ be any parameter of interest, $\hat{\theta}$ its estimate and $s(\hat{\theta})$ its asymptotic standard error. Consider the studentized statistic

$$t_n(\theta) = \frac{\hat{\theta} - \theta}{s(\hat{\theta})}. \tag{6.16}$$

Theorem 6.8.1 $t_n(\theta) \to_d N(0,1)$

Proof. By (6.15)

$$\begin{aligned} t_n(\theta) &= \frac{\hat{\theta} - \theta}{s(\hat{\theta})} \\ &= \frac{\sqrt{n}\left(\hat{\theta} - \theta\right)}{\sqrt{\hat{V}_\theta}} \\ &\to_d \frac{N(0, V_\theta)}{\sqrt{V_\theta}} \\ &= N(0,1) \end{aligned}$$

■

Thus the asymptotic distribution of the t-ratio $t_n(\theta)$ is the standard normal. Since the standard normal distribution does not depend on the parameters, we say that $t_n(\theta)$ is **asymptotically pivotal**. In special cases (such as the normal regression model, see Section X), the statistic t_n has an exact t distribution, and is therefore exactly free of unknowns. In this case, we say that t_n is an **exactly pivotal** statistic. In general, however, pivotal statistics are unavailable and so we must rely on asymptotically pivotal statistics.

A simple null and composite hypothesis takes the form

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0 \end{aligned}$$

where θ_0 is some pre-specified value, and $\theta = h(\beta)$ is some function of the parameter vector. (For example, θ could be a single element of β).

The standard test for H_0 against H_1 is the t-statistic (or studentized statistic)

$$t_n = t_n(\theta_0) = \frac{\hat{\theta} - \theta_0}{s(\hat{\theta})}.$$

Under H_0, $t_n \to_d N(0,1)$. Let $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. That is, if $Z \sim N(0,1)$, then $P(Z > z_{\alpha/2}) = \alpha/2$ and $P(|Z| > z_{\alpha/2}) = \alpha$. For example, $z_{.025} = 1.96$ and $z_{.05} = 1.645$. A test of asymptotic significance α rejects H_0 if $|t_n| > z_{\alpha/2}$. Otherwise the test does not reject, or "accepts" H_0. This is because

$$\begin{aligned} P\left(\text{reject } H_0 \mid H_0 \text{ true}\right) &= P\left(|t_n| > z_{\alpha/2} \mid \theta = \theta_0\right) \\ &\to P\left(|Z| > z_{\alpha/2}\right) = \alpha. \end{aligned}$$

The rejection/acceptance dichotomy is associated with the Neyman-Pearson approach to hypothesis testing.

An alternative approach, associated with Fisher, is to report an asymptotic p-value. The asymptotic p-value for the above statistic is constructed as follows. Define the tail probability, or asymptotic p-value function

$$p(t) = P\left(|Z| > |t|\right) = 2\left(1 - \Phi(|t|)\right).$$

Then the asymptotic p-value of the statistic t_n is

$$p_n = p(t_n).$$

If the p-value p_n is small (close to zero) then the evidence against H_0 is strong. In a sense, p-values and hypothesis tests are equivalent since $p_n < \alpha$ if and only if $|t_n| > z_{\alpha/2}$. Thus an equivalent statement of a Neyman-Pearson test is to reject at the $\alpha\%$ level if and only if $p_n < \alpha$. The p-value is more general, however, in that the reader is allowed to pick the level of significance α, in contrast to Neyman-Pearson rejection/acceptance reporting where the researcher picks the level.

Another helpful observation is that the p-value function has simply made a unit-free transformation of the test statistic. That is, under H_0, $p_n \to_d U[0,1]$, so the "unusualness" of the test statistic can be compared to the easy-to-understand uniform distribution, regardless of the complication of the distribution of the original test statistic. To see this fact, note that the asymptotic distribution of $|t_n|$ is $F(x) = 1 - p(x)$. Thus

$$\begin{aligned} P\left(1 - p_n \leq u\right) &= P\left(1 - p(t_n) \leq u\right) \\ &= P\left(F(t_n) \leq u\right) \\ &= P\left(|t_n| \leq F^{-1}(u)\right) \\ &\to F\left(F^{-1}(u)\right) = u, \end{aligned}$$

establishing that $1 - p_n \to_d U[0,1]$, from which it follows that $p_n \to_d U[0,1]$.

6.9 Confidence Intervals

A confidence interval C_n is an interval estimate of θ, and is a function of the data and hence is random. It is designed to cover θ with high probability. Either $\theta \in C_n$ or $\theta \notin C_n$. The coverage probability is $P(\theta \in C_n)$.

We typically cannot calculate the exact coverage probability $P(\theta \in C_n)$. However we often can calculate the asymptotic coverage probability $\lim_{n\to\infty} P(\theta \in C_n)$. We say that C_n has asymptotic $(1-\alpha)\%$ coverage for θ if $P(\theta \in C_n) \to 1-\alpha$ as $n \to \infty$.

A good method for construction of a confidence interval is the collection of parameter values which are not rejected by a statistical test. The t-test of the previous setion rejects $H_0 : \theta_0 = \theta$ if $|t_n(\theta)| > z_{\alpha/2}$ where $t_n(\theta)$ is the t-statistic (6.16) and $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. A confidence interval is then constructed as the values of θ for which this test does not reject:

$$\begin{aligned} C_n &= \left\{\theta : |t_n(\theta)| \leq z_{\alpha/2}\right\} \\ &= \left\{\theta : -z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{s(\hat{\theta})} \leq z_{\alpha/2}\right\} \\ &= \left[\hat{\theta} - z_{\alpha/2}s(\hat{\theta}), \quad \hat{\theta} + z_{\alpha/2}s(\hat{\theta})\right]. \end{aligned} \tag{6.17}$$

While there is no hard-and-fast guideline for choosing the coverage probability $1-\alpha$, the most common professional choice is 95%, or $\alpha = .05$. This corresponds to selecting the confidence interval $\left[\hat{\theta} \pm 1.96s(\hat{\theta})\right] \approx \left[\hat{\theta} \pm 2s(\hat{\theta})\right]$. Thus values of θ within two standard errors of the estimated $\hat{\theta}$ are considered "reasonable" candidates for the true value θ, and values of θ outside two standard errors of the estimated $\hat{\theta}$ are considered unlikely or unreasonable candidates for the true value.

The interval has been constructed so that as $n \to \infty$,

$$P\left(\theta \in C_n\right) = P\left(|t_n(\theta)| \leq z_{\alpha/2}\right) \to P\left(|Z| \leq z_{\alpha/2}\right) = 1-\alpha.$$

and C_n is an asymptotic $(1-\alpha)\%$ confidence interval.

6.10 Wald Tests

Sometimes $\theta = h(\beta)$ is a $q \times 1$ vector, and it is desired to test the joint restrictions simultaneously. In this case the t-statistic approach does not work. We have the null and alternative

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0. \end{aligned}$$

The natural estimate of θ is $\hat{\theta} = h(\hat{\beta})$ and has asymptotic covariance matrix estimate

$$\hat{V}_\theta = \hat{H}_\beta' \hat{V} \hat{H}_\beta$$

where

$$\hat{H}_\beta = \frac{\partial}{\partial \beta} h(\hat{\beta}).$$

The Wald statistic for H_0 against H_1 is

$$\begin{aligned} W_n &= n\left(\hat{\theta} - \theta_0\right)' \hat{V}_\theta^{-1} \left(\hat{\theta} - \theta_0\right) \\ &= n\left(h(\hat{\beta}) - \theta_0\right)' \left(\hat{H}_\beta' \hat{V} \hat{H}_\beta\right)^{-1} \left(h(\hat{\beta}) - \theta_0\right). \end{aligned} \tag{6.18}$$

When h is a linear function of β, $h(\beta) = R'\beta$, then the Wald statistic takes the form

$$W_n = n\left(R'\hat{\beta} - \theta_0\right)'\left(R'\hat{V}R\right)^{-1}\left(R'\hat{\beta} - \theta_0\right).$$

The delta method (6.15) showed that $\sqrt{n}\left(\hat{\theta} - \theta\right) \to_d Z \sim N(0, V_\theta)$, and Theorem 6.5.1 showed that $\hat{V} \to_p V$. Furthermore, $H_\beta(\beta)$ is a continuous function of β, so by the continuous mapping theorem, $H_\beta(\hat{\beta}) \to_p H_\beta$. Thus $\hat{V}_\theta = \hat{H}_\beta'\hat{V}\hat{H}_\beta \to_p H_\beta' V H_\beta = V_\theta > 0$ if H_β has full rank q. Hence

$$W_n = n\left(\hat{\theta} - \theta_0\right)'\hat{V}_\theta^{-1}\left(\hat{\theta} - \theta_0\right) \to_d Z'V_\theta^{-1}Z = \chi^2_q,$$

by Theorem A.8.2. We have established:

Theorem 6.10.1 *Under H_0 and Assumption 6.3.1, if rank$(H_\beta) = q$, then $W_n \to_d \chi^2_q$, a chi-square random variable with q degrees of freedom.*

An asymptotic Wald test rejects H_0 in favor of H_1 if W_n exceeds $\chi^2_q(\alpha)$, the upper-α quantile of the χ^2_q distribution. For example, $\chi^2_1(.05) = 3.84 = z^2_{.025}$. The Wald test fails to reject if W_n is less than $\chi^2_q(\alpha)$. The asymptotic p-value for W_n is $p_n = p(W_n)$, where $p(x) = P\left(\chi^2_q \geq x\right)$ is the tail probability function of the χ^2_q distribution. As before, the test rejects at the $\alpha\%$ level iff $p_n < \alpha$, and p_n is asymptotically $U[0,1]$ under H_0.

6.11 F Tests

Take the linear model

$$Y = X_1\beta_1 + X_2\beta_2 + e$$

where X_1 is $n \times k_1$ and X_2 is $n \times k_2$ and $k + 1 = k_1 + k_2$. The null hypothesis is

$$H_0 : \beta_2 = 0.$$

In this case, $\theta = \beta_2$, and there are $q = k_2$ restrictions. Also $h(\beta) = R'\beta$ is linear with $R = \begin{pmatrix} 0 \\ I \end{pmatrix}$ a selector matrix. We know that the Wald statistic takes the form

$$\begin{aligned} W_n &= n\hat{\theta}'\hat{V}_\theta^{-1}\hat{\theta} \\ &= n\hat{\beta}_2'\left(R'\hat{V}R\right)^{-1}\hat{\beta}_2. \end{aligned}$$

What we will show in this section is that if $\hat{V}$ is replaced with $\hat{V}^0 = \hat{\sigma}^2\left(n^{-1}X'X\right)^{-1}$, the covariance matrix estimator valid under homoskedasticity, then the Wald statistic can be written in the form

$$W_n = n\left(\frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2}\right) \tag{6.19}$$

where

$$\tilde{\sigma}^2 = \frac{1}{n}\tilde{e}'\tilde{e}, \qquad \tilde{e} = Y - X_1\tilde{\beta}_1, \qquad \tilde{\beta}_1 = \left(X_1'X_1\right)^{-1}X_1'Y$$

are from OLS of Y on X_1, and

$$\hat{\sigma}^2 = \frac{1}{n}\hat{e}'\hat{e}, \qquad \hat{e} = Y - X\hat{\beta}, \qquad \hat{\beta} = \left(X'X\right)^{-1} X'Y$$

are from OLS of Y on $X = (X_1, X_2)$.

The elegant feature about (6.19) is that it is directly computable from the standard output from two simple OLS regressions, as the sum of squared errors is a typical output from statistical packages. This statistic is typically reported as an "F-statistic" which is defined as

$$F = \frac{n-k}{n}\frac{W_n}{k_2} = \frac{\left(\tilde{\sigma}^2 - \hat{\sigma}^2\right)/k_2}{\hat{\sigma}^2/(n-k)}.$$

While it should be emphasized that equality (6.19) only holds if $\hat{V}^0 = \hat{\sigma}^2\left(n^{-1}X'X\right)^{-1}$, still this formula often finds good use in reading applied papers. Because of this connection we call (6.19) the F form of the Wald statistic.

We now derive expression (6.19). First, note that partitioned matrix inversion (2.2)

$$R'\left(X'X\right)^{-1}R = R'\begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1} R = \left(X_2'M_1X_2\right)^{-1}$$

where $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$. Thus

$$\left(R'\hat{V}^0R\right)^{-1} = \hat{\sigma}^{-2}n^{-1}\left(R'\left(X'X\right)^{-1}R\right)^{-1} = \hat{\sigma}^{-2}n^{-1}\left(X_2'M_1X_2\right)$$

and

$$\begin{aligned} W_n &= n\hat{\beta}_2'\left(R'\hat{V}^0R\right)^{-1}\hat{\beta}_2 \\ &= \frac{\hat{\beta}_2'\left(X_2'M_1X_2\right)\hat{\beta}_2}{\hat{\sigma}^2}. \end{aligned}$$

To simplify this expression further, note that if we regress Y on X_1 alone, the residual is $\tilde{e} = M_1Y$. Now consider the residual regression of $\tilde{e}$ on $\tilde{X}_2 = M_1X_2$. By the FWL theorem, $\tilde{e} = \tilde{X}_2\hat{\beta}_2 + \hat{e}$ and $\tilde{X}_2'\hat{e} = 0$. Thus

$$\begin{aligned} \tilde{e}'\tilde{e} &= \left(\tilde{X}_2\hat{\beta}_2 + \hat{e}\right)'\left(\tilde{X}_2\hat{\beta}_2 + \hat{e}\right) \\ &= \hat{\beta}_2'\tilde{X}_2'\tilde{X}_2\hat{\beta}_2 + \hat{e}'\hat{e} \\ &= \hat{\beta}_2'X_2'M_1X_2\hat{\beta}_2 + \hat{e}'\hat{e}, \end{aligned}$$

or alternatively,

$$\hat{\beta}_2'X_2'M_1X_2\hat{\beta}_2 = \tilde{e}'\tilde{e} - \hat{e}'\hat{e}.$$

Also, since

$$\hat{\sigma}^2 = n^{-1}\hat{e}'\hat{e}$$

we conclude that

$$W_n = n\left(\frac{\tilde{e}'\tilde{e} - \hat{e}'\hat{e}}{\hat{e}'\hat{e}}\right) = n\left(\frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2}\right),$$

as claimed.

In many statistical packages, when an OLS regression is reported, an "F statistic" is reported. This is

$$F = \frac{\left(\tilde{\sigma}_y^2 - \hat{\sigma}^2\right)/(k-1)}{\hat{\sigma}^2/(n-k)}.$$

where

$$\tilde{\sigma}_y^2 = \frac{1}{n}(y - \overline{y})'(y - \overline{y})$$

is the sample variance of y_i, equivalently the residual variance from an intercept-only model. This special F statistic is testing the hypothesis that *all* slope coefficients (other than the intercept) are zero. This was a popular statistic in the early days of econometric reporting, when sample sizes were very small and researchers wanted to know if there was "any explanatory power" to their regression. This is rarely an issue today, as sample sizes are typically sufficiently large that this F statistic is highly "significant". While there are special cases where this F statistic is useful, these cases are atypical.

6.12 Normal Regression Model

As an alternative to asymptotic distribution theory, there is an exact distribution theory available for the normal linear regression model, introduced in Section 4.3. The modelling assumption that the error e_i is independent of x_i and $N(0, \sigma^2)$ can be be used to calculate a set of exact distribution results.

In particular, under the normality assumption the error vector e is independent of X and has distribution $N\left(0, I_n\sigma^2\right)$. Since linear functions of normals are also normal, this implies that conditional on X

$$\begin{pmatrix} \hat{\beta} - \beta \\ \hat{e} \end{pmatrix} = \begin{pmatrix} (X'X)^{-1}X' \\ M \end{pmatrix} e \sim N\left(0, \begin{pmatrix} \sigma^2 (X'X)^{-1} & 0 \\ 0 & \sigma^2 M \end{pmatrix}\right)$$

where $M = I - X(X'X)^{-1}X'$. Since uncorrelated normal variables are independent, it follows that $\hat{\beta}$ is independent of any function of the OLS residuals, including the estimated error variance s^2.

The spectral decomposition of M yields

$$M = H \begin{bmatrix} I_{n-k-1} & 0 \\ 0 & 0 \end{bmatrix} H'$$

(see equation (2.4)) where $H'H = I_n$. Let $u = \sigma^{-1}H'e \sim N(0, H'H) \sim N(0, I_n)$. Then

$$\begin{aligned} \frac{(n-k)s^2}{\sigma^2} &= \frac{1}{\sigma^2}\hat{e}'\hat{e} \\ &= \frac{1}{\sigma^2}e'Me \\ &= \frac{1}{\sigma^2}e'H \begin{bmatrix} I_{n-k} & 0 \\ 0 & 0 \end{bmatrix} H'e \\ &= u' \begin{bmatrix} I_{n-k} & 0 \\ 0 & 0 \end{bmatrix} u \\ &\sim \chi^2_{n-k}, \end{aligned}$$

a chi-square distribution with $n-k$ degrees of freedom. Furthermore, if standard errors are calculated using the homoskedastic formula (6.10)

$$\frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{s\sqrt{\left[(X'X)^{-1}\right]_{jj}}} \sim \frac{N\left(0, \sigma^2\left[(X'X)^{-1}\right]_{jj}\right)}{\sqrt{\frac{\sigma^2}{n-k}\chi^2_{n-k}}\sqrt{\left[(X'X)^{-1}\right]_{jj}}} = \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-k}}{n-k}}} \sim t_{n-k}$$

a t distribution with $n-k$ degrees of freedom.

We summarize these findings

Theorem 6.12.1 *If e_i is independent of x_i and distributed $N(0,\sigma^2)$, and standard errors are calculated using the homoskedastic formula (6.10) then*

- $\hat{\beta} \sim N\left(0, \sigma^2 (X'X)^{-1}\right)$
- $\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}$,
- $\frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t_{n-k}$

In Theorem 6.3.1 and Theorem 6.8.1 we showed that in large samples, $\hat{\beta}$ and t are approximately normally distributed. In contrast, Theorem 6.12.1 shows that under the strong assumption of normality, $\hat{\beta}$ has an exact normal distribution and t has an exact t distribution. As inference (confidence intervals) are based on the t-ratio, the notable distinction is between the $N(0,1)$ and t_{n-k} distributions. The critical values are quite close if $n-k \geq 30$, so as a practical matter it does not matter which distribution is used. (Unless the sample size is unreasonably small.)

Now let us partition $\beta = (\beta_1, \beta_2)$ and consider tests of the linear restriction

$$\begin{aligned} H_0 &: \beta_2 = 0 \\ H_1 &: \beta_2 \neq 0 \end{aligned}$$

In the context of parametric models, a good testing procedure is based on the likelihood ratio statistic, which is twice the difference in the log-likelihood function evaluated under the null and alternative hypotheses. The estimator under the alternative is the unrestricted estimator $(\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2)$ discussed above. The Gaussian log-likelihood at these estimates is

$$\begin{aligned} L_n(\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2) &= -\frac{n}{2}\log\left(2\pi\hat{\sigma}^2\right) - \frac{1}{2\hat{\sigma}^2}\hat{e}'\hat{e} \\ &= -\frac{n}{2}\log\left(\hat{\sigma}^2\right) - \frac{n}{2}\log(2\pi) - \frac{n}{2}. \end{aligned}$$

The MLE of the model under the null hypothesis is $(\tilde{\beta}_1, 0, \tilde{\sigma}^2)$ where $\tilde{\beta}_1$ is the OLS estimate from a regression of y_i on x_{1i} only, with residual variance $\tilde{\sigma}^2$. The log-likelihood of this model is

$$L_n(\tilde{\beta}_1, 0, \tilde{\sigma}^2) = -\frac{n}{2}\log\left(\tilde{\sigma}^2\right) - \frac{n}{2}\log(2\pi) - \frac{n}{2}.$$

The LR statistic for H_0 is

$$\begin{aligned} LR &= 2\left(L_n(\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2) - L_n(\tilde{\beta}_1, 0, \tilde{\sigma}^2)\right) \\ &= n\left(\log\left(\tilde{\sigma}^2\right) - \log\left(\hat{\sigma}^2\right)\right) \\ &= n\log\left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right). \end{aligned}$$

By a first-order Taylor series approximation

$$LR = n \log\left(1 + \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} - 1\right) \simeq n\left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} - 1\right) = W_n.$$

the F statistic.

6.13 Problems with Tests of NonLinear Hypotheses

While the t and Wald tests work well when the hypothesis is a linear restriction on β, they can work quite poorly when the restrictions are nonlinear. This can be seen by a simple example introduced by Lafontaine and White (1986). Take the model

$$\begin{aligned} y_i &= \beta + e_i \\ e_i &\sim N(0, \sigma^2) \end{aligned}$$

and consider the hypothesis

$$H_0 : \beta = 1.$$

Let $\hat{\beta}$ and $\hat{\sigma}^2$ be the sample mean and variance of y_i. Then the standard Wald test for H_0 is

$$W_n = n\frac{\left(\hat{\beta} - 1\right)^2}{\hat{\sigma}^2}.$$

Now notice that H_0 is equivalent to the hypothesis

$$H_0(s) : \beta^s = 1$$

for any positive integer s. Letting $h(\beta) = \beta^s$, and noting $H_\beta = s\beta^{s-1}$, we find that the standard Wald test for $H_0(s)$ is

$$W_n(s) = n\frac{\left(\hat{\beta}^s - 1\right)^2}{\hat{\sigma}^2 s^2 \hat{\beta}^{2s-2}}.$$

While the hypothesis $\beta^s = 1$ is unaffected by the choice of s, the statistic $W_n(s)$ varies with s. This is an unfortunate feature of the Wald statistic.

To demonstrate this effect, we have plotted in Figure 6.4 the Wald statistic $W_n(s)$ as a function of s, setting $n/\sigma^2 = 10$. The increasing solid line is for the case $\hat{\beta} = 0.8$. The decreasing dashed line is for the case $\hat{\beta} = 1.7$. It is easy to see that in each case there are values of s for which the test statistic is significant relative to asymptotic critical values, while there are other values of s for which test test statistic is insignificant. This is distressing since the choice of s seems arbitrary and irrelevant to the actual hypothesis.

Our first-order asymptotic theory is not useful to help pick s, as $W_n(s) \to_d \chi_1^2$ under H_0 for any s. This is a context where **Monte Carlo simulation** can be quite useful as a tool to study and compare the exact distributions statistical procedures in finite samples. The method uses random simulation to create an artificial dataset to apply the statistical tools of interest. This produces random draws from the sampling distribution of interest. Through repetition, features of this distribution can be calculated.

In the present context of the Wald statistic, one feature of importance is the Type I error of the test using the asymptotic 5% critical value 3.84 – the probability of a false rejection,

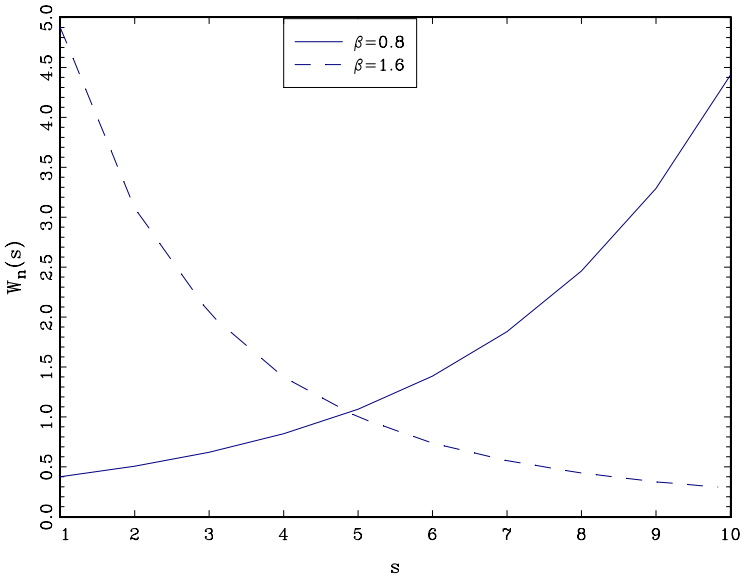

Figure 6.4: Wald Statistic as a function of s

$P\left(W_n(s) > 3.84 \mid \beta = 1\right)$. Given the simplicity of the model, this probability depends only on s, n, and σ^2. In Table 2.1 we report the results of a Monte Carlo simulation where we vary these three parameters. The value of s is varied from 1 to 10, n is varied among 20, 100 and 500, and σ is varied among 1 and 3. Table 4.1 reports the simulation estimate of the Type I error probability from 50,000 random samples. Each row of the table corresponds to a different value of s – and thus corresponds to a particular choice of test statistic. The second through seventh columns contain the Type I error probabilities for different combinations of n and σ. These probabilities are calculated as the percentage of the 50,000 simulated Wald statistics $W_n(s)$ which are larger than 3.84. The null hypothesis $\beta^s = 1$ is true, so these probabilities are Type I error.

To interpret the table, remember that the ideal Type I error probability is 5% (.05) with deviations indicating+ distortion. Typically, Type I error rates between 3% and 8% are considered reasonable. Error rates avove 10% are considered excessive. Rates above 20% are unexceptable. When comparing statistical procedures, we compare the rates row by row, looking for tests for which rate rejection rates are close to 5%, and rarely fall outside of the 3%-8% range. For this particular example, the only test which meets this criterion is the conventional $W_n = W_n(1)$ test. Any other choice of s leads to a test with unacceptable Type I error probabilities.

In Table 4.1 you can also see the impact of variation in sample size. In each case, the Type I error probability improves towards 5% as the sample size n increases. There is, however, no magic choice of n for which all tests perform uniformly well. Test performance deteriorates as s increases, which is not surprising given the dependence of $W_n(s)$ on s as shown in Figure 6.4.

Table 4.1
Type I error Probability of Asymptotic 5% $W_n(s)$ Test

	$\sigma = 1$			$\sigma = 3$		
s	$n = 20$	$n = 100$	$n = 500$	$n = 20$	$n = 100$	$n = 500$
1	.06	.05	.05	.07	.05	.05
2	.08	.06	.05	.15	.08	.06
3	.10	.06	.05	.21	.12	.07
4	.13	.07	.06	.25	.15	.08
5	.15	.08	.06	.28	.18	.10
6	.17	.09	.06	.30	.20	.11
7	.19	.10	.06	.31	.22	.13
8	.20	.12	.07	.33	.24	.14
9	.22	.13	.07	.34	.25	.15
10	.23	.14	.08	.35	.26	.16

Note: Rejection frequencies from 50,000 simulated random samples

In this example it is not surprising that the choice $s = 1$ yields the best test statistic. Other choices are arbitrary and would not be used in practice. While this is clear in this particular example, in other examples natural choices are not always obvious and the best choices may in fact appear counter-intuitive at first.

This point can be illustrated through another example. Take the model

$$\begin{aligned} y_i &= \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + e_i \\ E(x_i e_i) &= 0 \end{aligned} \tag{6.20}$$

and the hypothesis

$$H_0 : \frac{\beta_1}{\beta_2} = r$$

where r is a known constant. Equivalently, define $\theta = \beta_1/\beta_2$, so the hypothesis can be stated as $H_0 : \theta = r$.

Let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ be the least-squares estimates of (6.20), let $\hat{V}$ be an estimate of the asymptotic variance matrix for $\hat{\beta}$ and set $\hat{\theta} = \hat{\beta}_1/\hat{\beta}_2$. Define

$$\hat{H}_1 = \begin{pmatrix} 0 \\ \dfrac{1}{\hat{\beta}_2} \\ -\dfrac{\hat{\beta}_1}{\hat{\beta}_2^2} \end{pmatrix}$$

so that the standard error for $\hat{\theta}$ is $s(\hat{\theta}) = \left(n^{-1}\hat{H}_1'\hat{V}\hat{H}_1\right)^{1/2}$. In this case a t-statistic for H_0 is

$$t_{1n} = \frac{\left(\frac{\hat{\beta}_1}{\hat{\beta}_2} - r\right)}{s(\hat{\theta})}.$$

An alternative statistic can be constructed through reformulating the null hypothesis as

$$H_0 : \beta_1 - r\beta_2 = 0.$$

A t-statistic based on this formulation of the hypothesis is

$$t_{2n} = \frac{\left(\hat{\beta}_1 - r\hat{\beta}_2\right)^2}{\left(n^{-1}H_2\hat{V}H_2\right)^{1/2}}.$$

where

$$H_2 = \begin{pmatrix} 0 \\ 1 \\ -r \end{pmatrix}.$$

To compare t_{1n} and t_{2n} we perform another simple Monte Carlo simulation. We let x_{1i} and x_{2i} be mutually independent $N(0,1)$ variables, e_i be an independent $N(0,\sigma^2)$ draw with $\sigma = 3$, and normalize $\beta_0 = 0$ and $\beta_1 = 1$. This leaves β_2 as a free parameter, along with sample size n. We vary β_2 among .1, .25, .50, .75, and 1.0 and n among 100 and 500.

Table 4.2
Type I error Probability of Asymptotic 5% t-tests

	$n = 100$				$n = 500$			
	$P(t_n < -1.645)$		$P(t_n > 1.645)$		$P(t_n < -1.645)$		$P(t_n > 1.645)$	
β_2	t_{1n}	t_{2n}	t_{1n}	t_{2n}	t_{1n}	t_{2n}	t_{1n}	t_{2n}
.10	.47	.06	.00	.06	.28	.05	.00	.05
.25	.26	.06	.00	.06	.15	.05	.00	.05
.50	.15	.06	.00	.06	.10	.05	.00	.05
.75	.12	.06	.00	.06	.09	.05	.00	.05
1.00	.10	.06	.00	.06	.07	.05	.02	.05

The one-sided Type I error probabilities $P(t_n < -1.645)$ and $P(t_n > 1.645)$ are calculated from 50,000 simulated samples. The results are presented in Table 4.2. Ideally, the entries in the table should be 0.05. However, the rejection rates for the t_{1n} statistic diverge greatly from this value, especially for small values of β_2. The left tail probabilities $P(t_{1n} < -1.645)$ greatly exceed 5%, while the right tail probabilities $P(t_{1n} > 1.645)$ are close to zero in most cases. In contrast, the rejection rates for the linear t_{2n} statistic are invariant to the value of β_2, and are close to the ideal 5% rate for both sample sizes. The implication of Table 4.2 is that the two t-ratios have dramatically different sampling behavior.

The common message from both examples is that Wald statistics are sensitive to the algebraic formulation of the null hypothesis. In all cases, if the hypothesis can be expressed as a linear restriction on the model parameters, this formulation should be used. If no linear formulation is feasible, then the "most linear" formulation should be selected, and alternatives to asymptotic critical values should be considered. It is also prudent to consider alternative tests to the Wald statistic, such as the GMM distance statistic which will be presented in Section 9.7.

6.14 Monte Carlo Simulation

In the previous section we introduced the method of Monte Carlo simulation to illustrate the small sample problems with tests of nonlinear hypotheses. In this section we describe the method in more detail.

Recall, our data consist of observations (y_i, x_i) which are random draws from a population distribution F. Let θ be a parameter and let $T_n = T_n(y_1, x_1..., y_n, x_n, \theta)$ be a statistic of interest, for example an estimator $\hat{\theta}$ or a t-statistic $(\hat{\theta} - \theta)/s(\hat{\theta})$. The exact distribution of T_n is

$$G_n(x, F) = P\left(T_n \leq x \mid F\right).$$

While the asymptotic distribution of T_n might be known, the exact (finite sample) distribution G_n is generally unknown.

Monte Carlo simulation uses numerical simulation to compute $G_n(x, F)$ for selected choices of F. This is useful to investigate the performance of the statistic T_n in reasonable situations and sample sizes. The basic idea is that for any given F, the distribution function $G_n(x, F)$ can be calculated numerically through simulation. The name Monte Carlo derives from the famous Mediterranean gambling resort, where games of chance are played.

The method of Monte Carlo is quite simple to describe. The researcher chooses F (the distribution of the data) and the sample size n. A "true" value of θ is implied by this choice, or equivalently the value θ is selected directly by the researcher, which implies restrictions on F.

Then the following experiment is conducted

- n independent random pairs $\left(y_i^*, x_i^*\right)$, $i = 1, ..., n$, are drawn from the distribution F using the computer's random number generator.
- The statistic $T_n = T_n(y_1^*, x_n^*..., y_n^*, x_n^*, \theta)$ is calculated on this pseudo data.

For step 1, most computer packages have built-in procedures for generating $U[0, 1]$ and $N(0, 1)$ random numbers, and from these most random variables can be constructed. (For example, a chi-square can be generated by sums of squares of normals.)

For step 2, it is important that the statistic be evaluated at the "true" value of θ corresponding to the choice of F.

The above experiment creates one random draw from the distribution $G_n(x, F)$. This is one observation from an unknown distribution. Clearly, from one observation very little can be said. So the researcher repeats the experiment B times, where B is a large number. Typically, we set $B = 1000$ or $B = 5000$. We will discuss this choice later.

Notationally, let the $b'th$ experiment result in the draw T_{nb}, $b = 1, ..., B$. These results are stored. They constitute a random sample of size B from the distribution of $G_n(x, F) = P\left(T_{nb} \leq x\right) = P\left(T_n \leq x \mid F\right)$.

From a random sample, we can estimate any feature of interest using (typically) a method of moments estimator. For example:

Suppose we are interested in the bias, mean-squared error (MSE), or variance of the distribution of $\hat{\theta} - \theta$. We then set $T_n = \hat{\theta} - \theta$, run the above experiment, and calculate

$$\begin{aligned}
\widehat{Bias(\hat{\theta})} &= \frac{1}{B}\sum_{b=1}^{B} T_{nb} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b - \theta \\
\widehat{MSE(\hat{\theta})} &= \frac{1}{B}\sum_{b=1}^{B} \left(T_{nb}\right)^2 \\
\widehat{Var(\hat{\theta})} &= \widehat{MSE(\hat{\theta})} - \left(\widehat{Bias(\hat{\theta})}\right)^2
\end{aligned}$$

Suppose we are interested in the Type I error associated with an asymptotic 5% two-sided t-test. We would then set $T_n = \left|\hat{\theta} - \theta\right| / s(\hat{\theta})$ and calculate

$$\hat{P} = \frac{1}{B}\sum_{b=1}^{B} 1\left(T_{nb} \geq 1.96\right), \tag{6.21}$$

the percentage of the simulated t-ratios which exceed the asymptotic 5% critical value.

Suppose we are interested in the 5% and 95% quantile of $T_n = \hat{\theta}$. We then compute the 10% and 90% sample quantiles of the sample $\{T_{nb}\}$. The α% sample quantile is a number q_α such that α% of the sample are less than q_α. A simple way to compute sample quantiles is to sort the sample $\{T_{nb}\}$ from low to high. Then q_α is the N'th number in this ordered sequence, where $N = (B+1)\alpha$. It is therefore convenient to pick B so that N is an integer. For example, if we set $B = 999$, then the 5% sample quantile is 50'th sorted value and the 95% sample quantile is the 950'th sorted value.

The typical purpose of a Monte Carlo simulation is to investigate the performance of a statistical procedure (estimator or test) in realistic settings. Generally, the performance will depend on n and F. In many cases, an estimator or test may perform wonderfully for some values, and poorly for others. It is therefore useful to conduct a variety of experiments, for a selection of choices of n and F.

As discussed above, the researcher must select the number of experiments, B. Often this is called the number of **replications**. Quite simply, a larger B results in more precise estimates of the features of interest of G_n, but requires more computational time. In practice, therefore, the choice of B is often guided by the computational demands of the statistical procedure. Since the results of a Monte Carlo experiment are estimates computed from a random sample of size B, and therefore it is straightforward to calculate standard errors for any quantity of interest. If the standard error is too large to make a reliable inference, then B will have to be increased.

In particular, it is simple to make inferences about rejection probabilities from statistical tests, such as the percentage estimate reported in (6.21). The random variable $1\left(T_{nb} \geq 1.96\right)$ is iid Bernoulli, equalling 1 with probability $P = E1\left(T_{nb} \geq 1.96\right)$. The average (6.21) is therefore an unbiased estimator of P with standard error $s\left(\hat{P}\right) = \sqrt{P\left(1-P\right)/B}$. As P is unknown, this may be approximated by replacing P with $\hat{P}$ or with an hypothesized value. For example, if we are assessing an asymptotic 5% test, then we can set $s\left(\hat{P}\right) = \sqrt{(.05)\left(.95\right)/B} \simeq .22/\sqrt{B}$. Hence the standard errors for $B = 100$, 1000, and 5000, are, respectively, $s\left(\hat{P}\right) = .022$, .007, and .003.

6.15 Estimating a Wage Equation

We again return to our wage equation. We now expand the sample all non-military wage earners, and estimate a multivariate regression. Again our dependent variable is the natural log of wages, and our regressors include years of education, potential work experience, experience squared, and dummy variable indicators for the following: married, female, union member, immigrant, and hispanic. We separately estimate equations for white and non-whites.

For the dependent variable we use the natural log of wages, so that coefficients may be interpreted as semi-elasticities. We us the sample of wage earners from the March 2004 Current Population Survey, excluding military. For regressors we include years of education, potential work experience, experience squared, and dummy variable indicators for the following: married, female, union member, immigrant, hispanic, and non-white. Furthermore, we included a dummy

variable for state of residence (including the District of Columbia, this adds 50 regressors). The available sample is 18,808 so the parameter estimates are quite precise and reported in Table 4.1, excluding the coefficients on the state dummy variables.

Table 4.1
OLS Estimates of Linear Equation for Log(Wage)

	$\hat{\beta}$	$s(\hat{\beta})$
Intercept	1.027	.032
Education	.101	.002
Experience	.033	.001
Experience^2	−.00057	.00002
Married	.102	.008
Female	−.232	.007
Union Member	.097	.010
Immigrant	−.121	.013
Hispanic	−.102	.014
Non-White	−.070	.010
$\hat{\sigma}$	.4877	
Sample Size	18,808	
R^2	.34	

One question is whether or not the state dummy variables are relevant. Computing the Wald statistic (6.18) that the state coefficients are jointly zero, we find $W_n = 550$. Alternatively, re-estimating the model with the 50 state dummies excluded, the restricted standard deviation estimate is $\tilde{\sigma} = .4945$. The F form of the Wald statistic (6.19) is

$$W_n = n\left(1 - \frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\right) = 18,808\left(1 - \frac{.4877^2}{.4945^2}\right) = 515.$$

Notice that the two statistics are close, but not equal. Using either statistic the hypothesis is easily rejected, as the 1% critical value for the χ^2_{50} distribution is 76.

Another interesting question which can be addressed from these estimates is the maximal impact of experience on mean wages. Ignoring the other coefficients, we can write this effect as

$$\log(Wage) = \beta_2 Experience + \beta_3 Experience^2 + \cdots$$

Our question is: At which level of experience θ do workers achieve the highest wage? In this quadratic model, if $\beta_2 > 0$ and $\beta_3 < 0$ the solution is

$$\theta = -\frac{\beta_2}{2\beta_3}.$$

From Table 4.1 we find the point estimate

$$\hat{\theta} = -\frac{\hat{\beta}_2}{2\hat{\beta}_3} = 28.69.$$

Using the Delta Method, we can calculate a standard error of $s(\hat{\theta}) = .40$, implying a 95% confidence interval of [27.9, 29.5].

However, this is a poor choice, as the coverage probability of this confidence interval is one minus the Type I error of the hypothesis test based on the t-test. In the previous section we discovered that such t-tests had very poor Type I error rates. Instead, we found better Type I error rates by reformulating the hypothesis as a linear restriction. These t-statistics take the form

$$t_n(\theta) = \frac{\hat{\beta}_2 + 2\hat{\beta}_3\theta}{\left(h_\theta' \hat{V} h_\theta\right)^{1/2}}$$

where

$$h_\theta = \begin{pmatrix} -1 \\ 2\theta \end{pmatrix}$$

and $\hat{V}$ is the covariance matrix for $(\hat{\beta}_2\ \hat{\beta}_3)$.

In the present context we are interested in forming a confidence interval, not testing a hypothesis, so we have to go one step further. Our desired confidence interval will be the set of parameter values θ which are not rejected by the hypothesis test. This is the set of θ such that $|t_n(\theta)| \leq 1.96$. Since $t_n(\theta)$ is a non-linear function of θ, there is not a simple expression for this set, but it can be found numerically quite easily. This set is [27.0, 29.5]. Notice that the upper end of the confidence interval is the same as that from the delta method, but the lower end is substantially lower.

6.16 Exercises

For exercises 1-4, the following definition is used. In the model $Y = X\beta + e$, the least-squares estimate of β subject to the restriction $h(\beta) = 0$ is

$$\begin{aligned} \tilde{\beta} &= \underset{h(\beta)=0}{\operatorname{argmin}} S_n(\beta) \\ S_n(\beta) &= (Y - X\beta)'(Y - X\beta). \end{aligned}$$

That is, $\tilde{\beta}$ minimizes the sum of squared errors $S_n(\beta)$ over all β such that the restriction holds.

1. In the model $Y = X_1\beta_1 + X_2\beta_2 + e$, show that the least-squares estimate of $\beta = (\beta_1, \beta_2)$ subject to the constraint that $\beta_2 = 0$ is the OLS regression of Y on X_1.

2. In the model $Y = X_1\beta_1 + X_2\beta_2 + e$, show that the least-squares estimate of $\beta = (\beta_1, \beta_2)$, subject to the constraint that $\beta_1 = c$ (where c is some given vector) is simply the OLS regression of $Y - X_1 c$ on X_2.

3. In the model $Y = X_1\beta_1 + X_2\beta_2 + e$, find the least-squares estimate of $\beta = (\beta_1, \beta_2)$, subject to the constraint that $\beta_1 = -\beta_2$.

4. Take the model $Y = X\beta + e$ with the restriction $R'\beta = r$ where R is a known $k \times s$ matrix, r is a known $s \times 1$ vector, $0 < s < k$, and $rank(R) = s$. Explain why $\tilde{\beta}$ solves the minimization of the Lagrangian

$$L(\beta, \lambda) = \frac{1}{2} S_n(\beta) + \lambda' \left(R'\beta - r\right)$$

where λ is $s \times 1$.

   (a) Show that the solution is

$$\begin{aligned} \tilde{\beta} &= \hat{\beta} - \left(X'X\right)^{-1} R \left[R' \left(X'X\right)^{-1} R\right]^{-1} \left(R'\hat{\beta} - r\right) \\ \hat{\lambda} &= \left[R' \left(X'X\right)^{-1} R\right]^{-1} \left(R'\hat{\beta} - r\right) \end{aligned}$$

   where

$$\hat{\beta} = \left(X'X\right)^{-1} X'Y$$

   is the unconstrained OLS estimator.

   (b) Verify that $R'\tilde{\beta} = r$.

   (c) Show that if $R'\beta = r$ is true, then

$$\tilde{\beta} - \beta = \left(I_k - \left(X'X\right)^{-1} R \left[R' \left(X'X\right)^{-1} R\right]^{-1} R'\right) \left(X'X\right)^{-1} X'e.$$

   (d) Under the standard assumptions plus $R'\beta = r$, find the asymptotic distribution of $\sqrt{n}\left(\tilde{\beta} - \beta\right)$ as $n \to \infty$.

   (e) Find an appropriate formula to calculate standard errors for the elements of $\tilde{\beta}$.

5. You have two independent samples (Y_1, X_1) and (Y_2, X_2) which satisfy $Y_1 = X_1\beta_1 + e_1$ and $Y_2 = X_2\beta_2 + e_2$, where $E(x_{1i}e_{i1}) = 0$ and $E(x_{2i}e_{2i}) = 0$, and both X_1 and X_2 have k columns. Let $\hat{\beta}_1$ and $\hat{\beta}_2$ be the OLS estimates of β_1 and β_2. For simplicity, you may assume that both samples have the same number of observations n.

   (a) Find the asymptotic distribution of $\sqrt{n}\left(\left(\hat{\beta}_2 - \hat{\beta}_1\right) - (\beta_2 - \beta_1)\right)$ as $n \to \infty$.

   (b) Find an appropriate test statistic for $H_0 : \beta_2 = \beta_1$.

   (c) Find the asymptotic distribution of this statistic under H_0.

6. The model is

$$\begin{aligned} y_i &= x_i'\beta + e_i \\ E(x_i e_i) &= 0 \\ \Omega &= E\left(x_i x_i' e_i^2\right). \end{aligned}$$

   (a) Find the method of moments estimators $(\hat{\beta}, \hat{\Omega})$ for (β, Ω).

   (b) In this model, are $(\hat{\beta}, \hat{\Omega})$ efficient estimators of (β, Ω)?

   (c) If so, in what sense are they efficient?

7. Take the model $y_i = x_{1i}'\beta_1 + x_{2i}'\beta_2 + e_i$ with $Ex_ie_i = 0$. Suppose that β_1 is estimated by regressing y_i on x_{1i} only. Find the probability limit of this estimator. In general, is it consistent for β_1? If not, under what conditions is this estimator consistent for β_1?

8. Verify that equation (6.13) equals (6.14) as claimed in Section 6.6.

9. Prove that if an additional regressor X_{k+1} is added to X, Theil's adjusted $\overline{R}^2$ increases if and only if $|t_{k+1}| > 1$, where $t_{k+1} = \hat{\beta}_{k+1}/s(\hat{\beta}_{k+1})$ is the t-ratio for $\hat{\beta}_{k+1}$ and

$$s(\hat{\beta}_{k+1}) = \left(s^2[(X'X)^{-1}]_{k+1,k+1}\right)^{1/2}$$

   is the homoskedasticity-formula standard error.

10. Let Y be $n \times 1$, X be $n \times k$ (rank k). $Y = X\beta + e$ with $E(x_ie_i) = 0$. Define the *ridge regression* estimator

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i x_i' + \lambda I_k\right)^{-1} \left(\sum_{i=1}^{n} x_i y_i\right)$$

    where $\lambda > 0$ is a fixed constant. Find the probability limit of $\hat{\beta}$ as $n \to \infty$. Is $\hat{\beta}$ consistent for β?

11. Of the variables (y_i^*, y_i, x_i) only the pair (y_i, x_i) are observed. In this case, we say that y_i^* is a *latent* variable. Suppose

$$\begin{aligned} y_i^* &= x_i'\beta + e_i \\ E(x_i e_i) &= 0 \\ y_i &= y_i^* + u_i \end{aligned}$$

where u_i is a measurement error satisfying

$$\begin{aligned} E(x_i u_i) &= 0 \\ E(y_i^* u_i) &= 0 \end{aligned}$$

Let $\hat{\beta}$ denote the OLS coefficient from the regression of y_i on x_i.

(a) Is β the coefficient from the linear projection of y_i on x_i?

(b) Is $\hat{\beta}$ consistent for β as $n \to \infty$?

(c) Find the asymptotic distribution of $\sqrt{n}\left(\hat{\beta} - \beta\right)$ as $n \to \infty$.

12. The data set *invest.dat* contains data on 565 U.S. firms extracted from Compustat for the year 1987. The variables, in order, are

- I_i Investment to Capital Ratio (multiplied by 100).
- Q_i Total Market Value to Asset Ratio (Tobin's Q).
- C_i Cash Flow to Asset Ratio.
- D_i Long Term Debt to Asset Ratio.

The flow variables are annual sums for 1987. The stock variables are beginning of year.

(a) Estimate a linear regression of I_i on the other variables. Calculate appropriate standard errors.

(b) Calculate asymptotic confidence intervals for the coefficients.

(c) This regression is related to Tobin's q theory of investment, which suggests that investment should be predicted solely by Q_i. Thus the coefficient on Q_i should be positive and the others should be zero. Test the joint hypothesis that the coefficients on C_i and D_i are zero. Test the hypothesis that the coefficient on Q_i is zero. Are the results consistent with the predictions of the theory?

(d) Now try a non-linear (quadratic) specification. Regress I_i on Q_i, C_i, D_i, Q_i^2, C_i^2, D_i^2, $Q_i C_i$, $Q_i D_i$, $C_i D_i$. Test the joint hypothesis that the six interaction and quadratic coefficients are zero.

13. In a paper in 1963, Marc Nerlove analyzed a cost function for 145 American electric companies. (The problem is discussed in Example 8.3 of Greene, section 1.7 of Hayashi, and the empirical exercise in Chapter 1 of Hayashi). The data file *nerlov.dat* contains his data. The variables are described on page 77 of Hayashi. Nerlov was interested in estimating a *cost function*: $TC = f(Q, PL, PF, PK)$.

(a) First estimate an unrestricted Cobb-Douglass specification

$$\ln TC_i = \beta_1 + \beta_2 \ln Q_i + \beta_3 \ln PL_i + \beta_4 \ln PK_i + \beta_5 \ln PF_i + e_i. \tag{6.22}$$

Report parameter estimates and standard errors. You should obtain the same OLS estimates as in Hayashi's equation (1.7.7), but your standard errors may differ.

(b) Using a Wald statistic, test the hypothesis $H_0 : \beta_3 + \beta_4 + \beta_5 = 1$.

(c) Estimate (6.22) by least-squares imposing this restriction by substitution. Report your parameter estimates and standard errors.

(d) Estimate (6.22) subject to $\beta_3 + \beta_4 + \beta_5 = 1$ using the restricted least-squares estimator from problem 4. Do you obtain the same estimates as in part (c)?

.

Chapter 7

Additional Regression Topics

7.1 Generalized Least Squares

In the projection model, we know that the least-squares estimator is semi-parametrically efficient for the projection coefficient. However, in the linear regression model

$$\begin{aligned} y_i &= x_i'\beta + e_i \\ E(e_i \mid x_i) &= 0, \end{aligned}$$

the least-squares estimator is inefficient. The theory of Chamberlain (1987) can be used to show that in this model the semiparametric efficiency bound is obtained by the **Generalized Least Squares** (GLS) estimator

$$\tilde{\beta} = \left(X'D^{-1}X\right)^{-1}\left(X'D^{-1}Y\right) \tag{7.1}$$

where $D = diag\{\sigma_1^2, ..., \sigma_n^2\}$ and $\sigma_i^2 = \sigma^2(x_i) = E\left(e_i^2 \mid x_i\right)$. The GLS estimator is sometimes called the Aitken estimator. The GLS estimator (7.1) infeasible since the matrix D is unknown. A feasible GLS (FGLS) estimator replaces the unknown D with an estimate $\hat{D} = diag\{\hat{\sigma}_1^2, ..., \hat{\sigma}_n^2\}$. We now discuss this estimation problem.

Suppose that we model the conditional variance using the parametric form

$$\begin{aligned} \sigma_i^2 &= \alpha_0 + z_{1i}'\alpha_1 \\ &= \alpha' z_i, \end{aligned}$$

where z_{1i} is some $q \times 1$ function of x_i. Typically, z_{1i} are squares (and perhaps levels) of some (or all) elements of x_i. Often the functional form is kept simple for parsimony.

Let $\eta_i = e_i^2$. Then

$$E(\eta_i \mid x_i) = \alpha_0 + z_{1i}'\alpha_1$$

and we have the regression equation

$$\begin{aligned} \eta_i &= \alpha_0 + z_{1i}'\alpha_1 + \xi_i \\ E(\xi_i \mid x_i) &= 0. \end{aligned} \tag{7.2}$$

The error ξ_i in this regression error ξ_i is generally heteroskedastic and has the conditional variance

$$\begin{aligned} Var(\xi_i \mid x_i) &= Var\left(e_i^2 \mid x_i\right) \\ &= E\left(\left(e_i^2 - E\left(e_i^2 \mid x_i\right)\right)^2 \mid x_i\right) \\ &= E\left(e_i^4 \mid x_i\right) - \left(E\left(e_i^2 \mid x_i\right)\right)^2. \end{aligned}$$

Suppose e_i (and thus η_i) were observed. Then we could estimate α by OLS:

$$\hat{\alpha} = \left(Z'Z\right)^{-1} Z'\eta \to_p \alpha$$

and

$$\sqrt{n}\left(\hat{\alpha} - \alpha\right) \to_d N\left(0, V_\alpha\right)$$

where

$$V_\alpha = \left(E\left(z_i z_i'\right)\right)^{-1} E\left(z_i z_i' \xi_i^2\right) \left(E\left(z_i z_i'\right)\right)^{-1}. \tag{7.3}$$

While e_i is not observed, we have the OLS residual $\hat{e}_i = y_i - x_i'\hat{\beta} = e_i - x_i'(\hat{\beta} - \beta)$. Thus

$$\begin{aligned} \hat{\eta} - \eta_i &= \hat{e}_i^2 - e_i^2 \\ &= -2e_i x_i' \left(\hat{\beta} - \beta\right) + (\hat{\beta} - \beta)' x_i x_i' (\hat{\beta} - \beta) \\ &= \phi_i, \end{aligned}$$

say. Note that

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n z_i \phi_i &= \frac{-2}{n} \sum_{i=1}^n z_i e_i x_i' \sqrt{n} \left(\hat{\beta} - \beta\right) + \frac{1}{n} \sum_{i=1}^n z_i (\hat{\beta} - \beta)' x_i x_i' (\hat{\beta} - \beta) \sqrt{n} \\ &\to_p 0 \end{aligned}$$

Let

$$\tilde{\alpha} = \left(Z'Z\right)^{-1} Z'\hat{\eta} \tag{7.4}$$

be from OLS regression of $\hat{\eta}_i$ on z_i. Then

$$\begin{aligned} \sqrt{n}\left(\tilde{\alpha} - \alpha\right) &= \sqrt{n}\left(\hat{\alpha} - \alpha\right) + \left(n^{-1} Z'Z\right)^{-1} n^{-1/2} Z'\phi \\ &\to_d N\left(0, V_\alpha\right) \end{aligned} \tag{7.5}$$

Thus the fact that η_i is replaced with $\hat{\eta}_i$ is asymptotically irrelevant. We may call (7.4) the *skedastic* regression, as it is estimating the conditional variance of the regression of y_i on x_i. We have shown that α is consistently estimated by a simple procedure, and hence we can estimate $\sigma_i^2 = z_i'\alpha$ by

$$\tilde{\sigma}_i^2 = \tilde{\alpha}' z_i. \tag{7.6}$$

Suppose that $\tilde{\sigma}_i^2 > 0$ for all i. Then set

$$\tilde{D} = diag\{\tilde{\sigma}_1^2, ..., \tilde{\sigma}_n^2\}$$

and

$$\tilde{\beta} = \left(X'\tilde{D}^{-1}X\right)^{-1} X'\tilde{D}^{-1}Y.$$

This is the feasible GLS, or FGLS, estimator of β. Since there is not a unique specification for the conditional variance the FGLS estimator is not unique, and will depend on the model (and estimation method) for the skedastic regression.

One typical problem with implementation of FGLS estimation is that in a linear regression specification, there is no guarantee that $\tilde{\sigma}_i^2 > 0$ for all i. If $\tilde{\sigma}_i^2 < 0$ for some i, then the FGLS estimator is not well defined. Furthermore, if $\tilde{\sigma}_i^2 \approx 0$ for some i, then the FGLS estimator will force the regression equation through the point (y_i, x_i), which is typically undesirable. This suggests

that there is a need to bound the estimated variances away from zero. A trimming rule might make sense:

$$\overline{\sigma}_i^2 = \max[\tilde{\sigma}_i^2, \underline{\sigma}^2]$$

for some $\underline{\sigma}^2 > 0$.

It is possible to show that if the skedastic regression is correctly specified, then FGLS is asymptotically equivalent to GLS, but the proof of this can be tricky. We just state the result without proof.

Theorem 7.1.1 *If the skedastic regression is correctly specified,*

$$\sqrt{n}\left(\tilde{\beta}_{GLS} - \tilde{\beta}_{FGLS}\right) \to_p 0,$$

and thus

$$\sqrt{n}\left(\tilde{\beta}_{FGLS} - \beta\right) \to_d N(0, V),$$

where

$$V = \left(E\left(\sigma_i^{-2} x_i x_i'\right)\right)^{-1}.$$

Examining the asymptotic distribution of Theorem 7.1.1, the natural estimator of the asymptotic variance of $\hat{\beta}$ is

$$\tilde{V}^0 = \left(\frac{1}{n}\sum_{i=1}^n \tilde{\sigma}_i^{-2} x_i x_i'\right)^{-1} = \left(\frac{1}{n} X' \tilde{D}^{-1} X\right)^{-1}.$$

which is consistent for V as $n \to \infty$. This estimator $\tilde{V}^0$ is appropriate when the skedastic regression (7.2) is correctly specified.

It may be the case that $\alpha' z_i$ is only an approximation to the true conditional variance $\sigma_i^2 = E(e_i^2 \mid x_i)$. In this case we interpret $\alpha' z_i$ as a linear projection of e_i^2 on z_i. $\tilde{\beta}$ should perhaps be called a quasi-FGLS estimator of β. Its asymptotic variance is not that given in Theorem 7.1.1. Instead,

$$V = \left(E\left(\left(\alpha' z_i\right)^{-1} x_i x_i'\right)\right)^{-1} \left(E\left(\left(\alpha' z_i\right)^{-2} \sigma_i^2 x_i x_i'\right)\right) \left(E\left(\left(\alpha' z_i\right)^{-1} x_i x_i'\right)\right)^{-1}.$$

V takes a sandwich form, similar to the covariance matrix of the OLS estimator. Unless $\sigma_i^2 = \alpha' z_i$, $\tilde{V}^0$ is inconsistent for V.

An appropriate solution is to use a White-type estimator in place of $\tilde{V}^0$. This may be written as

$$\begin{aligned}
\tilde{V} &= \left(\frac{1}{n}\sum_{i=1}^n \tilde{\sigma}_i^{-2} x_i x_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n \tilde{\sigma}_i^{-4} \hat{e}_i^2 x_i x_i'\right) \left(\frac{1}{n}\sum_{i=1}^n \tilde{\sigma}_i^{-2} x_i x_i'\right)^{-1} \\
&= n\left(X' \tilde{D}^{-1} X\right)^{-1} \left(X' \tilde{D}^{-1} \hat{D} \tilde{D}^{-1} X\right) \left(X' \tilde{D}^{-1} X\right)^{-1}
\end{aligned}$$

where $\hat{D} = diag\{\hat{e}_1^2, ..., \hat{e}_n^2\}$. This is an estimator which is robust to misspecification of the conditional variance, and was proposed by Cragg (*Journal of Econometrics*, 1992).

In the linear regression model, FGLS is asymptotically superior to OLS. Why then do we not exclusively estimate regression models by FGLS? This is a good question. There are three reasons.

First, FGLS estimation depends on specification and estimation of the skedastic regression. Since the form of the skedastic regression is unknown, and it may be estimated with considerable

error, the estimated conditional variances may contain more noise than information about the true conditional variances. In this case, FGLS will do worse than OLS in practice.

Second, individual estimated conditional variances may be negative, and this requires trimming to solve. This introduces an element of arbitrariness which is unsettling to empirical researchers.

Third, OLS is a more robust estimator of the parameter vector. It is consistent not only in the regression model, but also under the assumptions of linear projection. The GLS and FGLS estimators, on the other hand, require the assumption of a correct conditional mean. If the equation of interest is a linear projection, and not a conditional mean, then the OLS and FGLS estimators will converge in probability to different limits, as they will be estimating two different projections. And the FGLS probability limit will depend on the particular function selected for the skedastic regression. The point is that the efficiency gains from FGLS are built on the stronger assumption of a correct conditional mean, and the cost is a reduction of robustness to misspecificatio

7.2 Testing for Heteroskedasticity

The hypothesis of homoskedasticity is that $E\left(e_i^2 \mid x_i\right) = \sigma^2$, or equivalently that

$$H_0 : \alpha_1 = 0$$

in the regression (7.2). We may therefore test this hypothesis by the estimation (7.4) and constructing a Wald statistic.

This hypothesis does not imply that ξ_i is independent of x_i. Typically, however, we impose the stronger hypothesis and test the hypothesis that e_i is independent of x_i, in which case ξ_i is independent of x_i and the asymptotic variance (7.3) for $\tilde{\alpha}$ simplifies to

$$V_\alpha = \left(E\left(z_i z_i'\right)\right)^{-1} E\left(\xi_i^2\right). \tag{7.7}$$

Hence the standard test of H_0 is a classic F (or Wald) test for exclusion of all regressors from the skedastic regression (7.4). The asymptotic distribution (7.5) and the asymptotic variance (7.7) under independence show that this test has an asymptotic chi-square distribution.

Theorem 7.2.1 *Under H_0 and e_i independent of x_i, the Wald test of H_0 is asymptotically χ_q^2.*

Most tests for heteroskedasticity take this basic form. The main differences between popular "tests" is which transformations of x_i enter z_i. Motivated by the form of the asymptotic variance of the OLS estimator $\hat{\beta}$, White (1980) proposed that the test for heteroskedasticity be based on setting z_i to equal all non-redundant elements of x_i, its squares, and all cross-products. Breusch-Pagan (1979) proposed what might appear to be a distinct test, but the only difference is that they allowed for general choice of z_i, and replaced $E\left(\xi_i^2\right)$ with $2\sigma^4$ which holds when e_i is $N(0, \sigma^2)$. If this simplification is replaced by the standard formula (under independence of the error), the two tests coincide.

7.3 Forecast Intervals

In the linear regression model the conditional mean of y_i given $x_i = x$ is

$$m(x) = E\left(y_i \mid x_i = x\right) = x'\beta.$$

In some cases, we want to estimate $m(x)$ at a particular point x. Notice that this is a (linear) function of β. Letting $h(\beta) = x'\beta$ and $\theta = h(\beta)$, we see that $\hat{m}(x) = \hat{\theta} = x'\hat{\beta}$ and $H_\beta = x$, so $s(\hat{\theta}) = \sqrt{n^{-1}x'\hat{V}x}$. Thus an asymptotic 95% confidence interval for $m(x)$ is

$$\left[x'\hat{\beta} \pm 2\sqrt{n^{-1}x'\hat{V}x}\right].$$

It is interesting to observe that if this is viewed as a function of x, the width of the confidence set is dependent on x.

For a given value of $x_i = x$, we may want to forecast (guess) y_i out-of-sample. A reasonable rule is the conditional mean $m(x)$ as it is the mean-square-minimizing forecast. A point forecast is the estimated conditional mean $\hat{m}(x) = x'\hat{\beta}$. We would also like a measure of uncertainty for the forecast.

The forecast error is $\hat{e}_i = y_i - \hat{m}(x) = e_i - x'\left(\hat{\beta} - \beta\right)$. As the out-of-sample error e_i is independent of the in-sample estimate $\hat{\beta}$, this has variance

$$\begin{aligned} E\hat{e}_i^2 &= E\left(e_i^2 \mid x_i = x\right) + x'E\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)' x \\ &= \sigma^2(x) + n^{-1}x'Vx. \end{aligned}$$

Assuming $E\left(e_i^2 \mid x_i\right) = \sigma^2$, the natural estimate of this variance is $\hat{\sigma}^2 + n^{-1}x'\hat{V}x$, so a standard error for the forecast is $\hat{s}(x) = \sqrt{\hat{\sigma}^2 + n^{-1}x'\hat{V}x}$. Notice that this is different from the standard error for the conditional mean. If we have an estimate of the conditional variance function, e.g. $\tilde{\sigma}^2(x) = \tilde{\alpha}'z'$ from (7.6), then the forecast standard error is $\hat{s}(x) = \sqrt{\tilde{\sigma}^2(x) + n^{-1}x'\hat{V}x}$

It would appear natural to conclude that an asymptotic 95% forecast interval for y_i is

$$\left[x'\hat{\beta} \pm 2\hat{s}(x)\right],$$

but this turns out to be incorrect. In general, the validity of an asymptotic confidence interval is based on the asymptotic normality of the studentized ratio. In the present case, this would require the asymptotic normality of the ratio

$$\frac{e_i - x'\left(\hat{\beta} - \beta\right)}{\hat{s}(x)}.$$

But no such asymptotic approximation can be made. The only special exception is the case where e_i has the exact distribution $N(0, \sigma^2)$, which is generally invalid.

To get an accurate forecast interval, we need to estimate the conditional distribution of e_i given $x_i = x$, which is a much more difficult task. Given the difficulty, many applied forecasters focus on the simple approximate interval $\left[x'\hat{\beta} \pm 2\hat{s}(x)\right]$.

7.4 NonLinear Least Squares

In some cases we might use a parametric regression function $m(x, \theta) = E\left(y_i \mid x_i = x\right)$ which is a non-linear function of the parameters θ. We describe this setting as **non-linear regression**.

Examples of nonlinear regression functions include

$$\begin{aligned}
m(x,\theta) &= \theta_1 + \theta_2 \frac{x}{1+\theta_3 x} \\
m(x,\theta) &= \theta_1 + \theta_2 x^{\theta_3} \\
m(x,\theta) &= \theta_1 + \theta_2 \exp(\theta_3 x) \\
m(x,\theta) &= G(x'\theta),\ G \text{ known} \\
m(x,\theta) &= \theta_1 + \theta_2 x_1 + (\theta_3 + \theta_4 x_1)\, \Phi\left(\frac{x_2 - \theta_5}{\theta_6}\right) \\
m(x,\theta) &= \theta_1 + \theta_2 x + \theta_4 (x - \theta_3)\, 1(x > \theta_3) \\
m(x,\theta) &= (\theta_1 + \theta_2 x_1)\, 1(x_2 < \theta_3) + (\theta_4 + \theta_5 x_1)\, 1(x_2 > \theta_3)
\end{aligned}$$

In the first five examples, $m(x,\theta)$ is (generically) differentiable in the parameters θ. In the final two examples, m is not differentiable with respect to θ_3, which alters some of the analysis. When it exists, let

$$m_\theta(x,\theta) = \frac{\partial}{\partial \theta} m(x,\theta).$$

Nonlinear regression is frequently adopted because the functional form $m(x,\theta)$ is suggested by an economic model. In other cases, it is adopted as a flexible approximation to an unknown regression function.

The least squares estimator $\hat{\theta}$ minimizes the sum-of-squared-errors

$$S_n(\theta) = \sum_{i=1}^{n} (y_i - m(x_i,\theta))^2.$$

When the regression function is nonlinear, we call this the **nonlinear least squares** (NLLS) estimator. The NLLS residuals are $\hat{e}_i = y_i - m(x_i,\hat{\theta})$.

One motivation for the choice of NLLS as the estimation method is that the parameter θ is the solution to the population problem $\min_\theta E(y_i - m(x_i,\theta))^2$

Since sum-of-squared-errors function $S_n(\theta)$ is not quadratic, $\hat{\theta}$ must be found by numerical methods. See Appendix E. When $m(x,\theta)$ is differentiable, then the FOC for minimization are

$$0 = \sum_{i=1}^{n} m_\theta(x_i,\hat{\theta})\hat{e}_i. \tag{7.8}$$

Theorem 7.4.1 *If the model is identified and $m(x,\theta)$ is differentiable with respect to θ,*

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to_d N(0,V)$$

$$V = \left(E\left(m_{\theta i} m_{\theta i}'\right)\right)^{-1} \left(E\left(m_{\theta i} m_{\theta i}' e_i^2\right)\right) \left(E\left(m_{\theta i} m_{\theta i}'\right)\right)^{-1}$$

where $m_{\theta i} = m_\theta(x_i,\theta_0)$.

Sketch of Proof. First, it must be shown that $\hat{\theta} \to_p \theta_0$. This can be done using arguments for optimization estimators, but we won't cover that argument here. Since $\hat{\theta} \to_p \theta_0$, $\hat{\theta}$ is close to θ_0 for n large, so the minimization of $S_n(\theta)$ only needs to be examined for θ close to θ_0. Let

$$y_i^0 = e_i + m_{\theta i}'\theta_0.$$

For θ close to the true value θ_0, by a first-order Taylor series approximation,

$$m(x_i, \theta) \simeq m(x_i, \theta_0) + m'_{\theta i} (\theta - \theta_0).$$

Thus

$$\begin{aligned} y_i - m(x_i, \theta) &\simeq (e_i + m(x_i, \theta_0)) - \big(m(x_i, \theta_0) + m'_{\theta i} (\theta - \theta_0)\big) \\ &= e_i - m'_{\theta i} (\theta - \theta_0) \\ &= y_i^0 - m'_{\theta i} \theta. \end{aligned}$$

Hence the sum of squared errors function is

$$S_n(\theta) = \sum_{i=1}^n (y_i - m(x_i, \theta))^2 \simeq \sum_{i=1}^n \left(y_i^0 - m'_{\theta i}\theta\right)^2$$

and the right-hand-side is the SSE function for a linear regression of y_i^0 on $m_{\theta i}$. Thus the NLLS estimator $\hat{\theta}$ has the same asymptotic distribution as the (infeasible) OLS regression of y_i^0 on $m_{\theta i}$, which is that stated in the theorem. ■

Based on Theorem 7.4.1, an estimate of the asymptotic variance V is

$$\hat{V} = \left(\frac{1}{n}\sum_{i=1}^n \hat{m}_{\theta i}\hat{m}'_{\theta i}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n \hat{m}_{\theta i}\hat{m}'_{\theta i}\hat{e}_i^2\right) \left(\frac{1}{n}\sum_{i=1}^n \hat{m}_{\theta i}\hat{m}'_{\theta i}\right)^{-1}$$

where $\hat{m}_{\theta i} = m_\theta(x_i, \hat{\theta})$ and $\hat{e}_i = y_i - m(x_i, \hat{\theta})$.

Identification is often tricky in nonlinear regression models. Suppose that

$$m(x_i, \theta) = \beta_1' z_i + \beta_2' x_i(\gamma).$$

The model is linear when $\beta_2 = 0$, and this is often a useful hypothesis (sub-model) to consider. Thus we want to test

$$H_0 : \beta_2 = 0.$$

However, under H_0, the model is

$$y_i = \beta_1' z_i + \varepsilon_i$$

and both β_2 and γ have dropped out. This means that under H_0, γ is not identified. This renders the distribution theory presented in the previous section invalid. Thus when the truth is that $\beta_2 = 0$, the parameter estimates are not asymptotically normally distributed. Furthermore, tests of H_0 do not have asymptotic normal or chi-square distributions.

The asymptotic theory of such tests have been worked out by Andrews and Ploberger (1994) and B. Hansen (1996). In particular, Hansen shows how to use simulation (similar to the bootstrap) to construct the asymptotic critical values (or p-values) in a given application.

7.5 Least Absolute Deviations

We stated that a conventional goal in econometrics is estimation of impact of variation in x_i on the central tendency of y_i. We have discussed projections and conditional means, but these are not the only measures of central tendency. An alternative good measure is the conditional median.

To recall the definition and properties of the median, let Y be a continuous random variable. The median $\theta_0 = Med(Y)$ is the value such that $P(Y \leq \theta_0) = P(Y \geq \theta_0) = .5$. Two useful facts about the median are that

$$\theta_0 = \underset{\theta}{\operatorname{argmin}} E|Y - \theta| \tag{7.9}$$

and

$$E \operatorname{sgn}(Y - \theta_0) = 0$$

where

$$\operatorname{sgn}(u) = \begin{cases} 1 & \text{if } u \geq 0 \\ -1 & \text{if } u < 0 \end{cases}$$

is the sign function.

These facts definitions motivate three estimators of θ. The first definition is the $50th$ empirical quantile. The second is the value which minimizes $\frac{1}{n}\sum_{i=1}^{n}|y_i - \theta|$, and the third definition is the solution to the moment equation $\frac{1}{n}\sum_{i=1}^{n}\operatorname{sgn}(y_i - \theta)$. These distinctions are illusory, however, as these estimators are indeed identical.

Now let's consider the conditional median of Y given a random variable X. Let $m(x) = Med(Y \mid X = x)$ denote the conditional median of Y given $X = x$, and let $Med(Y \mid X) = m(X)$ be this function evaluated at the random variable X. The linear median regression model takes the form

$$\begin{aligned} y_i &= x_i'\beta + e_i \\ Med(e_i \mid x_i) &= 0 \end{aligned}$$

In this model, the linear function $Med(y_i \mid x_i = x) = x'\beta$ is the conditional median function, and the substantive assumption is that the median function is linear in x.

Conditional analogs of the facts about the median are

- $P(y_i \leq x'\beta_0 \mid x_i = x) = P(y_i > x'\beta \mid x_i = x) = .5$
- $E(\operatorname{sgn}(e_i) \mid x_i) = 0$
- $E(x_i \operatorname{sgn}(e_i)) = 0$
- $\beta_0 = \min_\beta E|y_i - x_i'\beta|$

These facts motivate the following estimator. Let

$$L_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - x_i'\beta\right|$$

be the average of absolute deviations. The **least absolute deviations** (LAD) estimator of β minimizes this function

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} L_n(\beta)$$

Equivalently, it is a solution to the moment condition

$$\frac{1}{n}\sum_{i=1}^{n} x_i \operatorname{sgn}\left(y_i - x_i'\hat{\beta}\right) = 0. \tag{7.10}$$

The LAD estimator has the asymptotic distribution

Theorem 7.5.1 $\sqrt{n}\left(\hat{\beta}-\beta_0\right) \to_d N(0,V)$, *where*

$$V = \frac{1}{4}\left(E\left(x_i x_i' f\left(0 \mid x_i\right)\right)\right)^{-1}\left(E x_i x_i'\right)\left(E\left(x_i x_i' f\left(0 \mid x_i\right)\right)\right)^{-1}$$

and $f\left(e \mid x\right)$ *is the conditional density of* e_i *given* $x_i = x$.

The variance of the asymptotic distribution inversely depends on $f\left(0 \mid x\right)$, the conditional density of the error at its median. When $f\left(0 \mid x\right)$ is large, then there are many innovations near to the median, and this improves estimation of the median. In the special case where the error is independent of x_i, then $f\left(0 \mid x\right) = f\left(0\right)$ and the asymptotic variance simplifies

$$V = \frac{\left(E x_i x_i'\right)^{-1}}{4f\left(0\right)^2} \tag{7.11}$$

This simplification is similar to the simplification of the asymptotic covariance of the OLS estimator under homoskedasticity.

Computation of standard error for LAD estimates typically is based on equation (7.11). The main difficulty is the estimation of $f(0)$, the height of the error density at its median. This can be done with kernel estimation techniques. See Chapter 16. While a complete proof of Theorem 7.5.1 is advanced, we provide a sketch here for completeness.

Proof of Theorem 7.5.1: Since $\operatorname{sgn}\left(a\right) = 1 - 2 \cdot 1\left(a \leq 0\right)$, (7.10) is equivalent to $\overline{g}_n(\hat{\beta}) = 0$, where $\overline{g}_n(\beta) = n^{-1}\sum_{i=1}^n g_i(\beta)$ and $g_i(\beta) = x_i\left(1 - 2 \cdot 1\left(y_i \leq x_i'\beta\right)\right)$. Let $g(\beta) = E g_i(\beta)$. We need three preliminary result. First, by the central limit theorem (Theorm 5.3.1)

$$\sqrt{n}\left(\overline{g}_n(\beta_0) - g(\beta_0)\right) = -n^{-1/2}\sum_{i=1}^n g_i(\beta_0) \to_d N(0, E x_i x_i')$$

since $E g_i(\beta_0) g_i(\beta_0)' = E x_i x_i'$. Second using the law of iterated expectations and the chain rule of differentiation,

$$\begin{aligned}
\frac{\partial}{\partial \beta'} g(\beta) &= \frac{\partial}{\partial \beta'} E x_i\left(1 - 2 \cdot 1\left(y_i \leq x_i'\beta\right)\right) \\
&= -2\frac{\partial}{\partial \beta'} E\left[x_i E\left(1\left(e_i \leq x_i'\beta - x_i'\beta_0\right) \mid x_i\right)\right] \\
&= -2\frac{\partial}{\partial \beta'} E\left[x_i \int_{-\infty}^{x_i'\beta - x_i'\beta_0} f\left(e \mid x_i\right) de\right] \\
&= -2E\left[x_i x_i' f\left(x_i'\beta - x_i'\beta_0 \mid x_i\right)\right]
\end{aligned}$$

so

$$\frac{\partial}{\partial \beta'} g(\beta_0) = -2E\left[x_i x_i' f\left(0 \mid x_i\right)\right].$$

Third, by a Taylor series expansion and the fact $g(\beta_0) = 0$

$$g(\hat{\beta}) \simeq \frac{\partial}{\partial \beta'} g(\beta_0)\left(\hat{\beta} - \beta_0\right).$$

Together

$$\begin{aligned}
\sqrt{n}\left(\hat{\beta}-\beta_0\right) &\simeq \left(\frac{\partial}{\partial\beta'}g(\beta_0)\right)^{-1}\sqrt{n}g(\hat{\beta}) \\
&= \left(-2E\left[x_i x_i' f\left(0 \mid x_i\right)\right]\right)^{-1}\sqrt{n}\left(g(\hat{\beta})-\overline{g}_n(\hat{\beta})\right) \\
&\simeq \frac{1}{2}\left(E\left[x_i x_i' f\left(0 \mid x_i\right)\right]\right)^{-1}\sqrt{n}\left(\overline{g}_n(\beta_0)-g(\beta_0)\right) \\
&\to_d \frac{1}{2}\left(E\left[x_i x_i' f\left(0 \mid x_i\right)\right]\right)^{-1} N(0, Ex_i x_i') \\
&= N(0,V).
\end{aligned}$$

The third line follows from an asymptotic empirical process argument. ■

7.6 Quantile Regression

The method of quantile regression has become quite popular in recent econometric practice. For $\tau \in [0,1]$ the τ'th quantile Q_τ of a random variable with distribution function $F(u)$ is defined as

$$Q_\tau = \inf\{u : F(u) \geq \tau\}$$

When $F(u)$ is continuous and strictly monotonic, then $F\left(Q_\tau\right) = \tau$, so you can think of the quantile as the inverse of the distribution function. The quantile Q_τ is the value such that τ (percent) of the mass of the distribution is less than Q_τ. The median is the special case $\tau = .5$.

The following alternative representation is useful. If the random variable U has τ'th quantile Q_τ, then

$$Q_\tau = \underset{\theta}{\operatorname{argmin}}\, E\rho_\tau\left(U-\theta\right). \tag{7.12}$$

where $\rho_\tau\left(q\right)$ is the piecewise linear function

$$\begin{aligned}
\rho_\tau\left(q\right) &= \begin{cases} -q\left(1-\tau\right) & q<0 \\ q\tau & q\geq 0 \end{cases} \\
&= q\left(\tau - 1\left(q<0\right)\right).
\end{aligned} \tag{7.13}$$

This generalizes representation (7.9) for the median to all quantiles.

For the random variables (y_i, x_i) with conditional distribution function $F\left(y \mid x\right)$ the conditional quantile function $q_\tau(x)$ is

$$Q_\tau(x) = \inf\left\{y : F\left(y \mid x\right) \geq \tau\right\}.$$

Again, when $F\left(y \mid x\right)$ is continuous and strictly monotonic in y, then $F\left(Q_\tau(x) \mid x\right) = \tau$. For fixed τ, the quantile regression function $q_\tau(x)$ describes how the τ'th quantile of the conditional distribution varies with the regressors.

As functions of x, the quantile regression functions can take any shape. However for computational convenience it is typical to assume that they are (approximately) linear in x (after suitable transformations). This linear specification assumes that $Q_\tau(x) = \beta_\tau' x$ where the coefficients β_τ vary across the quantiles τ. We then have the linear quantile regression model

$$y_i = x_i'\beta_\tau + e_i$$

where e_i is the error defined to be the difference between y_i and its τ'th conditional quantile $x_i'\beta_\tau$. By construction, the τ'th conditional quantile of e_i is zero, otherwise its properties are unspecified without further restrictions.

Given the representation (7.12), the quantile regression estimator $\hat{\beta}_\tau$ for β_τ solves the minimization problem

$$\hat{\beta}_\tau = \underset{\beta \in R^k}{\operatorname{argmin}} L_n^\tau(\beta)$$

where

$$L_n^\tau(\beta) = \frac{1}{n}\sum_{i=1}^n \rho_\tau\left(y_i - x_i'\beta\right)$$

and $\rho_\tau(q)$ is defined in (7.13).

Since the quanitle regression criterion function $L_n^\tau(\beta)$ does not have an algebraic solution, numerical methods are necessary for its minimization. Furthermore, since it has discontinuous derivatives, conventional Newton-type optimization methods are inappropriate. Fortunately, fast linear programming methods have been developed for this problem, and are widely available.

A asymptotic distribution theory for the quantile regression estimator can be derived using similar arguments as those for the LAD estimator in Theorem 7.5.1.

Theorem 7.6.1 $\sqrt{n}\left(\hat{\beta}_\tau - \beta_\tau\right) \to_d N(0, V_\tau)$, *where*

$$V_\tau = \tau(1-\tau)\left(E\left(x_i x_i' f\left(0 \mid x_i\right)\right)\right)^{-1}\left(E x_i x_i'\right)\left(E\left(x_i x_i' f\left(0 \mid x_i\right)\right)\right)^{-1}$$

and $f(e \mid x)$ *is the conditional density of* e_i *given* $x_i = x$.

In general, the asymptotic variance depends on the conditional density of the quantile regression error. When the error e_i is independent of x_i, then $f\left(0 \mid x_i\right) = f(0)$, the unconditional density of e_i at 0, and we have the simplification

$$V_\tau = \frac{\tau(1-\tau)}{f(0)^2}\left(E\left(x_i x_i'\right)\right)^{-1}.$$

A recent monograph on the details of quantile regression is Koenker (2005).

7.7 Testing for Omitted NonLinearity

If the goal is to estimate the conditional expectation $E\left(y_i \mid x_i\right)$, it is useful to have a general test of the adequacy of the specification.

One simple test for neglected nonlinearity is to add nonlinear functions of the regressors to the regression, and test their significance using a Wald test. Thus, if the model $y_i = x_i'\hat{\beta} + \hat{e}_i$ has been fit by OLS, let $z_i = h(x_i)$ denote functions of x_i which are not linear functions of x_i (perhaps squares of non-binary regressors) and then fit $y_i = x_i'\tilde{\beta} + z_i'\tilde{\gamma} + \tilde{e}_i$ by OLS, and form a Wald statistic for $\gamma = 0$.

Another popular approach is the RESET test proposed by Ramsey (1969). The null model is

$$y_i = x_i'\beta + \varepsilon_i$$

which is estimated by OLS, yielding predicted values $\hat{y}_i = x_i'\hat{\beta}$. Now let

$$z_i = \begin{pmatrix} \hat{y}_i^2 \\ \vdots \\ \hat{y}_i^m \end{pmatrix}$$

be an $(m-1)$-vector of powers of $\hat{y}_i$. Then run the auxiliary regression

$$y_i = x_i'\tilde{\beta} + z_i'\tilde{\gamma} + \tilde{e}_i \tag{7.14}$$

by OLS, and form the Wald statistic W_n for $\gamma = 0$. It is easy (although somewhat tedious) to show that under the null hypothesis, $W_n \to_d \chi^2_{m-1}$. Thus the null is rejected at the $\alpha\%$ level if W_n exceeds the upper $\alpha\%$ tail critical value of the χ^2_{m-1} distribution.

To implement the test, m must be selected in advance. Typically, small values such as $m = 2$, 3, or 4 seem to work best.

The RESET test appears to work well as a test of functional form against a wide range of smooth alternatives. It is particularly powerful at detecting *single-index* models of the form

$$y_i = G(x_i'\beta) + \varepsilon_i$$

where $G(\cdot)$ is a smooth "link" function. To see why this is the case, note that (7.14) may be written as

$$y_i = x_i'\tilde{\beta} + \left(x_i'\hat{\beta}\right)^2 \tilde{\gamma}_1 + \left(x_i'\hat{\beta}\right)^3 \tilde{\gamma}_2 + \cdots \left(x_i'\hat{\beta}\right)^m \tilde{\gamma}_{m-1} + \tilde{e}_i$$

which has essentially approximated $G(\cdot)$ by a m'th order polynomial.

7.8 Irrelevant Variables

In the model

$$\begin{aligned} y_i &= x_{1i}'\beta_1 + x_{2i}'\beta_2 + e_i \\ E\left(x_i e_i\right) &= 0, \end{aligned}$$

x_{2i} is "irrelevant" if β_1 is the parameter of interest and $\beta_2 = 0$. One estimator of β_1 is to regress y_i on x_{1i} alone, $\tilde{\beta}_1 = \left(X_1'X_1\right)^{-1}\left(X_1'Y\right)$. Another is to regress y_i on x_{1i} and x_{2i} jointly, yielding $(\hat{\beta}_1, \hat{\beta}_2)$. Under which conditions is $\hat{\beta}_1$ or $\tilde{\beta}_1$ superior?

It is easy to see that both estimators are consistent for β_1. However, they will (typically) have difference asymptotic variances.

The comparison between the two estimators is straightforward when the error is conditionally homoskedastic $E\left(e_i^2 \mid x_i\right) = \sigma^2$. In this case

$$\lim_{n\to\infty} nVar(\ \tilde{\beta}_1) = \left(Ex_{1i}x_{1i}'\right)^{-1}\sigma^2 = Q_{11}^{-1}\sigma^2,$$

say, and

$$\lim_{n\to\infty} nVar(\hat{\beta}_1) = \left(Ex_{1i}x_{1i}' - Ex_{1i}x_{2i}'\left(Ex_{2i}x_{2i}'\right)Ex_{2i}x_{1i}'\right)^{-1}\sigma^2 = \left(Q_{11} - Q_{12}Q_{22}^{-1}Q_{21}\right)^{-1}\sigma^2,$$

say. If $Q_{12} = 0$ (so the variables are orthogonal) then these two variance matrices equal, and the two estimators have equal asymptotic efficiency. Otherwise, since $Q_{12}Q_{22}^{-1}Q_{21} > 0$, then $Q_{11} > Q_{11} - Q_{12}Q_{22}^{-1}Q_{21}$, and consequently

$$Q_{11}^{-1}\sigma^2 < \left(Q_{11} - Q_{12}Q_{22}^{-1}Q_{21}\right)^{-1}\sigma^2.$$

This means that $\tilde{\beta}_1$ has a lower asymptotic variance matrix than $\hat{\beta}_1$. We conclude that the inclusion of irrelevant variable reduces estimation efficiency if these variables are correlated with the relevant variables.

For example, take the model $y_i = \beta_0 + \beta_1 x_i + e_i$ and suppose that $\beta_0 = 0$. Let $\hat{\beta}_1$ be the estimate of β_1 from the unconstrained model, and $\tilde{\beta}_1$ be the estimate under the constraint $\beta_0 = 0$. (The least-squares estimate with the intercept omitted.). Let $Ex_i = \mu$, and $E(x_i - \mu)^2 = \sigma_x^2$. Then under (6.7),

$$\lim_{n\to\infty} nVar(\tilde{\beta}_1) = \frac{\sigma^2}{\sigma_x^2 + \mu^2}$$

while

$$\lim_{n\to\infty} nVar(\hat{\beta}_1)^{-1} = \frac{\sigma^2}{\sigma_x^2}.$$

When $\mu \neq 0$, we see that $\tilde{\beta}_1$ has a lower asymptotic variance.

However, this result can be reversed when the error is conditionally heteroskedastic. In the absence of the homoskedasticity assumption, there is no clear ranking of the efficiency of the restricted estimator $\tilde{\beta}_1$ versus the unrestricted estimator.

7.9 Model Selection

In earlier sections we discussed the costs and benefits of inclusion/exclusion of variables. How does a researcher go about selecting an econometric specification, when economic theory does not provide complete guidance? This is the question of model selection. It is important that the model selection question be well-posed. For example, the question: "What is the right model for y?" is not well posed, because it does not make clear the conditioning set. In contrast, the question, "Which subset of $(x_1, ..., x_K)$ enters the regression function $E(y_i \mid x_{1i} = x_1, ..., x_{Ki} = x_K)$?" is well posed.

In many cases the problem of model selection can be reduced to the comparison of two nested models, as the larger problem can be written as a sequence of such comparisons. We thus consider the question of the inclusion of X_2 in the linear regression

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon,$$

where X_1 is $n \times k_1$ and X_2 is $n \times k_2$. This is equivalent to the comparison of the two models

$$\begin{aligned} \mathcal{M}_1 &: \quad Y = X_1\beta_1 + \varepsilon, \qquad\qquad E(\varepsilon \mid X_1, X_2) = 0 \\ \mathcal{M}_2 &: \quad Y = X_1\beta_1 + X_2\beta_2 + \varepsilon, \qquad E(\varepsilon \mid X_1, X_2) = 0. \end{aligned}$$

Note that $\mathcal{M}_1 \subset \mathcal{M}_2$. To be concrete, we say that $\mathcal{M}_2$ is true if $\beta_2 \neq 0$.

To fix notation, models 1 and 2 are estimated by OLS, with residual vectors $\hat{e}_1$ and $\hat{e}_2$, estimated variances $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, etc., respectively. To simplify some of the statistical discussion, we will on occasion use the homoskedasticity assumption $E\left(e_i^2 \mid x_{1i}, x_{2i}\right) = \sigma^2$.

A model selection procedure is a data-dependent rule which selects one of the two models. We can write this as $\widehat{\mathcal{M}}$. There are many possible desirable properties for a model selection procedure. One useful property is consistency, that it selects the true model with probability one if the sample is sufficiently large. A model selection procedure is consistent if

$$\begin{aligned} P\left(\widehat{\mathcal{M}} = \mathcal{M}_1 \mid \mathcal{M}_1\right) &\to 1 \\ P\left(\widehat{\mathcal{M}} = \mathcal{M}_2 \mid \mathcal{M}_2\right) &\to 1 \end{aligned}$$

However, this rule only makes sense when the true model is finite dimensional. If the truth is infinite dimensional, it is more appropriate to view model selection as determining the best finite sample approximation.

A common approach to model selection is to base the decision on a statistical test such as the Wald W_n. The model selection rule is as follows. For some critical level α, let c_α satisfy $P\left(\chi^2_{k_2} > c_\alpha\right)$. Then select $\mathcal{M}_1$ if $W_n \leq c_\alpha$, else select $\mathcal{M}_2$.

A major problem with this approach is that the critical level α is indeterminate. The reasoning which helps guide the choice of α in hypothesis testing (controlling Type I error) is not relevant for model selection. That is, if α is set to be a small number, then $P\left(\widehat{\mathcal{M}} = \mathcal{M}_1 \mid \mathcal{M}_1\right) \approx 1 - \alpha$ but $P\left(\widehat{\mathcal{M}} = \mathcal{M}_2 \mid \mathcal{M}_2\right)$ could vary dramatically, depending on the sample size, etc. Another problem is that if α is held fixed, then this model selection procedure is inconsistent, as $P\left(\widehat{\mathcal{M}} = \mathcal{M}_1 \mid \mathcal{M}_1\right) \to 1 - \alpha < 1$.

Another common approach to model selection is to to a selection criterion. One popular choice is the Akaike Information Criterion (AIC). The AIC for model m is

$$AIC_m = \log\left(\hat{\sigma}^2_m\right) + 2\frac{k_m}{n}. \tag{7.15}$$

where $\hat{\sigma}^2_m$ is the variance estimate for model m, and k_m is the number of coefficients in the model. The AIC can be derived as an estimate of the KullbackLeibler information distance $K(\mathcal{M}) = E\left(\log f(Y \mid X) - \log f(Y \mid X, \mathcal{M})\right)$ between the true density and the model density. The rule is to select $\mathcal{M}_1$ if $AIC_1 < AIC_2$, else select $\mathcal{M}_2$. AIC selection is inconsistent, as the rule tends to overfit. Indeed, since under $\mathcal{M}_1$,

$$LR_n = n\left(\log\hat{\sigma}^2_1 - \log\hat{\sigma}^2_2\right) \simeq W_n \to_d \chi^2_{k_2}, \tag{7.16}$$

then

$$\begin{aligned} P\left(\widehat{\mathcal{M}} = \mathcal{M}_1 \mid \mathcal{M}_1\right) &= P\left(AIC_1 < AIC_2 \mid \mathcal{M}_1\right) \\ &= P\left(\log(\hat{\sigma}^2_1) + 2\frac{k_1}{n} < \log(\hat{\sigma}^2_2) + 2\frac{k_1 + k_2}{n} \mid \mathcal{M}_1\right) \\ &= P\left(LR_n < 2k_2 \mid \mathcal{M}_1\right) \\ &\to P\left(\chi^2_{k_2} < 2k_2\right) < 1. \end{aligned}$$

While many modifications of the AIC have been proposed, the most popular to be one proposed by Schwarz, based on Bayesian arguments. His criterion, known as the BIC, is

$$BIC_m = \log\left(\hat{\sigma}^2_m\right) + \log(n)\frac{k_m}{n}. \tag{7.17}$$

Since $\log(n) > 2$ (if $n > 8$), the BIC places a larger penalty than the AIC on the number of estimated parameters and is more parsimonious.

In contrast to the AIC, BIC model selection is consistent. Indeed, since (7.16) holds under $\mathcal{M}_1$,

$$\frac{LR_n}{\log(n)} \to_p 0,$$

so

$$\begin{aligned} P\left(\widehat{\mathcal{M}} = \mathcal{M}_1 \mid \mathcal{M}_1\right) &= P\left(BIC_1 < BIC_2 \mid \mathcal{M}_1\right) \\ &= P\left(LR_n < \log(n) k_2 \mid \mathcal{M}_1\right) \\ &= P\left(\frac{LR_n}{\log(n)} < k_2 \mid \mathcal{M}_1\right) \\ &\to P\left(0 < k_2\right) = 1. \end{aligned}$$

Also under $\mathcal{M}_2$, one can show that

$$\frac{LR_n}{\log(n)} \to_p \infty,$$

thus

$$\begin{aligned} P\left(\widehat{\mathcal{M}} = \mathcal{M}_2 \mid \mathcal{M}_2\right) &= P\left(\frac{LR_n}{\log(n)} > k_2 \mid \mathcal{M}_2\right) \\ &\to 1. \end{aligned}$$

We have discussed model selection between two models. The methods extend readily to the issue of selection among multiple regressors. The general problem is the model

$$y_i = \beta_1 x_{1i} + \beta_2 x_{1i} + \cdots + \beta_K x_{Ki} + \varepsilon_i, \qquad E\left(\varepsilon_i \mid x_i\right) = 0$$

and the question is which subset of the coefficients are non-zero (equivalently, which regressors enter the regression).

There are two leading cases: ordered regressors and unordered.

In the ordered case, the models are

$$\begin{aligned} \mathcal{M}_1 &: \beta_1 \neq 0, \beta_2 = \beta_3 = \cdots = \beta_K = 0 \\ \mathcal{M}_2 &: \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 = \cdots = \beta_K = 0 \\ &\vdots \\ \mathcal{M}_K &: \beta_1 \neq 0, \beta_2 \neq 0, \ldots, \beta_K \neq 0. \end{aligned}$$

which are nested. The AIC selection criteria estimates the K models by OLS, stores the residual variance $\hat{\sigma}^2$ for each model, and then selects the model with the lowest AIC (7.15). Similarly for the BIC, selecting based on (7.17).

In the unordered case, a model consists of any possible subset of the regressors $\{x_{1i}, ..., x_{Ki}\}$, and the AIC or BIC in principle can be implemented by estimating all possible subset models. However, there are 2^K such models, which can be a very large number. For example, $2^{10} = 1024$, and $2^{20} = 1,048,576$. In the latter case, a full-blown implementation of the BIC selection criterion would seem computationally prohibitive.

7.10 Exercises

1. For any predictor $g(x_i)$ for y, the mean absolute error (MAE) is

$$E\left|y_i - g(x_i)\right|.$$

Show that the function $g(x)$ which minimizes the MAE is the conditional median $M(x)$.

2. Define

$$g(u) = \tau - 1\left(u < 0\right)$$

where $1\left(\cdot\right)$ is the indicator function (takes the value 1 if the argument is true, else equals zero). Let θ satisfy $Eg(Y_i - \theta) = 0$. Is θ a quantile of the distribution of Y_i?

3. Verify equation (7.12).

4. In the homoskedastic regression model $Y = X\beta + e$ with $E(e_i \mid x_i) = 0$ and $E(e_i^2 \mid x_i) = \sigma^2$, suppose $\hat{\beta}$ is the OLS estimate of β with covariance matrix $\hat{V}$, based on a sample of size n. Let $\hat{\sigma}^2$ be the estimate of σ^2. You wish to forecast an out-of-sample value of y given that $X = x$. Thus the available information is the sample (Y, X), the estimates $(\hat{\beta}, \hat{V}, \hat{\sigma}^2)$, the residuals $\hat{e}$, and the out-of-sample value of the regressors, x.

   (a) Find a point forecast of y.

   (b) Find an estimate of the variance of this forecast.

5. In a linear model

$$Y = X\beta + e, \quad E(e \mid X) = 0, \quad Var(e \mid X) = \sigma^2\Omega$$

with Ω known, the GLS estimator is

$$\tilde{\beta} = \left(X'\Omega^{-1}X\right)^{-1}\left(X'\Omega^{-1}Y\right).$$

the residual vector is $\hat{e} = Y - X\tilde{\beta}$, and an estimate of σ^2 is

$$s^2 = \frac{1}{n-k}\hat{e}'\Omega^{-1}\hat{e}.$$

   (a) Why is this a reasonable estimator for σ^2?

   (b) Prove that $\hat{e} = M_1 e$, where $M_1 = I - X\left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}$.

   (c) Prove that $M_1'\Omega^{-1}M_1 = \Omega^{-1} - \Omega^{-1}X\left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}$.

6. Let (y_i, x_i) be a random sample with $E(Y \mid X) = X\beta$. Consider the *Weighted Least Squares* (WLS) estimator of β

$$\tilde{\beta} = \left(X'WX\right)^{-1}\left(X'WY\right)$$

where $W = diag(w_1, ..., w_n)$ and $w_i = x_{ji}^{-2}$, where x_{ji} is one of the x_i.

   (a) In which contexts would $\tilde{\beta}$ be a good estimator?

   (b) Using your intuition, in which situations would you expect that $\tilde{\beta}$ would perform better than OLS?

7. Suppose that $y_i = g(x_i, \theta) + e_i$ with $E(e_i \mid x_i) = 0$, $\hat{\theta}$ is the NLLS estimator, and $\hat{V}$ is the estimate of $Var\left(\hat{\theta}\right)$. You are interested in the conditional mean function $E(y_i \mid x_i = x) = g(x)$ at some x. Find an asymptotic 95% confidence interval for $g(x)$.

8. The model is

$$\begin{aligned} y_i &= x_i\beta + e_i \\ E(e_i \mid x_i) &= 0 \end{aligned}$$

where $x_i \in R$. Consider the two estimators

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \\ \tilde{\beta} &= \frac{1}{n}\sum_{i=1}^n \frac{y_i}{x_i}. \end{aligned}$$

   (a) Under the stated assumptions, are both estimators consistent for β?

   (b) Are there conditions under which either estimator is efficient?

9. In Chapter 6, Exercise 13, you estimated a cost function on a cross-section of electric companies. The equation you estimated was

$$\ln TC_i = \beta_1 + \beta_2 \ln Q_i + \beta_3 \ln PL_i + \beta_4 \ln PK_i + \beta_5 \ln PF_i + e_i. \tag{7.18}$$

   (a) Following Nerlove, add the variable $(\ln Q_i)^2$ to the regression. Do so. Assess the merits of this new specification using (i) a hypothesis test; (ii) AIC criterion; (iii) BIC criterion. Do you agree with this modification?

   (b) Now try a non-linear specification. Consider model (7.18) plus the extra term $a_6 z_i$, where

$$z_i = \ln Q_i \left(1 + \exp\left(-\left(\ln Q_i - a_7\right)\right)\right)^{-1}.$$

   In addition, impose the restriction $a_3 + a_4 + a_5 = 1$. This model is called a smooth threshold model. For values of $\ln Q_i$ much below a_7, the variable $\ln Q_i$ has a regression slope of a_2. For values much above a_7, the regression slope is $a_2 + a_6$, and the model imposes a smooth transition between these regimes. The model is non-linear because of the parameter a_7.

   The model works best when a_7 is selected so that several values (in this example, at least 10 to 15) of $\ln Q_i$ are both below and above a_7. Examine the data and pick an appropriate range for a_7.

   (c) Estimate the model by non-linear least squares. I recommend the concentration method: Pick 10 (or more or you like) values of a_7 in this range. For each value of a_7, calculate z_i and estimate the model by OLS. Record the sum of squared errors, and find the value of a_7 for which the sum of squared errors is minimized.

   (d) Calculate standard errors for all the parameters $(a_1, ..., a_7)$.

10. The data file *cps78.dat* contains 550 observations on 20 variables taken from the May 1978 current population survey. Variables are listed in the file *cps78.pdf*. The goal of the exercise is to estimate a model for the log of earnings (variable LNWAGE) as a function of the conditioning variables.

    (a) Start by an OLS regression of LNWAGE on the other variables. Report coefficient estimates and standard errors.

    (b) Consider augmenting the model by squares and/or cross-products of the conditioning variables. Estimate your selected model and report the results.

    (c) Are there any variables which seem to be unimportant as a determinant of wages? You may re-estimate the model without these variables, if desired.

    (d) Test whether the error variance is different for men and women. Interpret.

    (e) Test whether the error variance is different for whites and nonwhites. Interpret.

    (f) Construct a model for the conditional variance. Estimate such a model, test for general heteroskedasticity and report the results.

    (g) Using this model for the conditional variance, re-estimate the model from part (c) using FGLS. Report the results.

    (h) Do the OLS and FGLS estimates differ greatly? Note any interesting differences.

    (i) Compare the estimated standard errors. Note any interesting differences.

Chapter 8

The Bootstrap

8.1 Definition of the Bootstrap

Let F denote a distribution function for the population of observations (y_i, x_i). Let

$$T_n = T_n(y_1, x_{1,}..., y_n, x_n, F)$$

be a statistic of interest, for example an estimator $\hat{\theta}$ or a t-statistic $\left(\hat{\theta} - \theta\right)/s(\hat{\theta})$. Note that we write T_n as possibly a function of F. For example, the t-statistic is a function of the parameter θ which itself is a function of F.

The exact CDF of T_n when the data are sampled from the distribution F is

$$G_n(x, F) = P(T_n \leq x \mid F)$$

In general, $G_n(x, F)$ depends on F, meaning that G changes as F changes.

Ideally, inference would be based on $G_n(x, F)$. This is generally impossible since F is unknown.

Asymptotic inference is based on approximating $G_n(x, F)$ with $G(x, F) = \lim_{n\to\infty} G_n(x, F)$. When $G(x, F) = G(x)$ does not depend on F, we say that T_n is asymptotically pivotal and use the distribution function $G(x)$ for inferential purposes.

In a seminal contribution, Efron (1979) proposed the bootstrap, which makes a different approximation. The unknown F is replaced by a consistent estimate F_n (one choice is discussed in the next section). Plugged into $G_n(x, F)$ we obtain

$$G_n^*(x) = G_n(x, F_n). \tag{8.1}$$

We call G_n^* the bootstrap distribution. Bootstrap inference is based on $G_n^*(x)$.

Let (y_i^*, x_i^*) denote random variables with the distribution F_n. A random sample from this distribution is call the bootstrap data. The statistic $T_n^* = T_n(y_1^*, x_{1,}^*..., y_n^*, x_n^*, F_n)$ constructed on this sample is a random variable with distribution G_n^*. That is, $P(T_n^* \leq x) = G_n^*(x)$. We call T_n^* the bootstrap statistic. The distribution of T_n^* is identical to that of T_n when the true CDF of F_n rather than F.

The bootstrap distribution is itself random, as it depends on the sample through the estimator F_n.

In the next sections we describe computation of the bootstrap distribution.

8.2 The Empirical Distribution Function

Recall that $F(y,x) = P(y_i \leq y, x_i \leq x) = E(1(y_i \leq y)\,1(x_i \leq x))$, where $1(\cdot)$ is the indicator function. This is a population moment. The method of moments estimator is the corresponding sample moment:

$$F_n(y,x) = \frac{1}{n}\sum_{i=1}^{n} 1(y_i \leq y)\,1(x_i \leq x). \tag{8.2}$$

$F_n(y,x)$ is called the empirical distribution function (EDF). F_n is a nonparametric estimate of F. Note that while F may be either discrete or continuous, F_n is by construction a step function.

The EDF is a consistent estimator of the CDF. To see this, note that for any (y,x), $1(y_i \leq y)\,1(x_i \leq x)$ is an iid random variable with expectation $F(y,x)$. Thus by the WLLN (Theorem 5.2.1), $F_n(y,x) \to_p F(y,x)$. Furthermore, by the CLT (Theorem 5.3.1),

$$\sqrt{n}\,(F_n(y,x) - F(y,x)) \to^d N(0, F(y,x)(1 - F(y,x))).$$

To see the effect of sample size on the EDF, in the Figure below, I have plotted the EDF and true CDF for three random samples of size $n = 25$, 50, and 100. The random draws are from the $N(0,1)$ distribution. For $n = 25$, the EDF is only a crude approximation to the CDF, but the approximation appears to improve for the large n. In general, as the sample size gets larger, the EDF step function gets uniformly close to the true CDF.

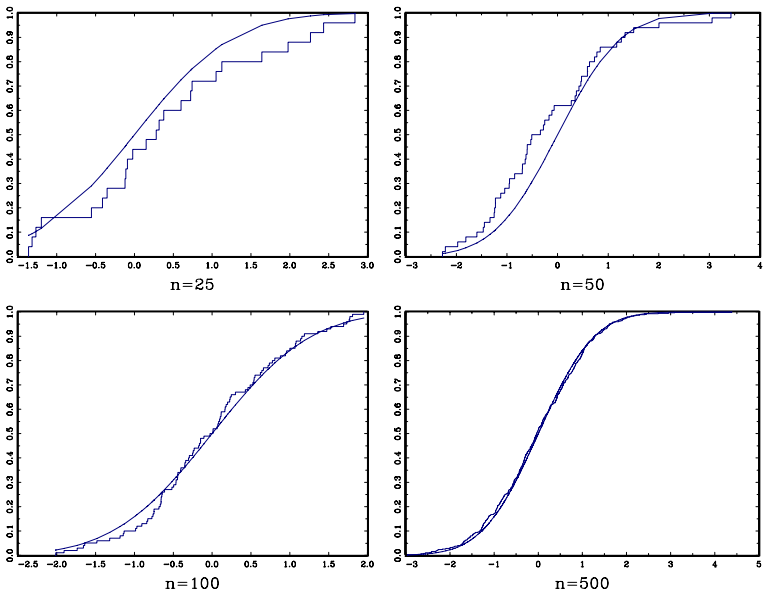

Figure 8.1: Empirical Distribution Functions

The EDF is a valid discrete probability distribution which puts probability mass $1/n$ at each pair (y_i, x_i), $i = 1, ..., n$. Notationally, it is helpful to think of a random pair (y_i^*, x_i^*) with the distribution F_n. That is,

$$P(y_i^* \leq y, x_i^* \leq x) = F_n(y,x).$$

We can easily calculate the moments of functions of (y_i^*, x_i^*) :

$$\begin{aligned} Eh\left(y_i^*, x_i^*\right)) &= \int h(y,x) dF_n(y,x) \\ &= \sum_{i=1}^{n} h\left(y_i, x_i\right) P\left(y_i^* = y_i, x_i^* = x_i\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} h\left(y_i, x_i\right), \end{aligned}$$

the empirical sample average.

8.3 Nonparametric Bootstrap

The **nonparametric bootstrap** is obtained when the bootstrap distribution (8.1) is defined using the EDF (8.2) as the estimate F_n of F.

Since the EDF F_n is a multinomial (with n support points), in principle the distribution G_n^* could be calculated by direct methods. However, as there are 2^n possible samples $\{(y_1^*, x_1^*), ..., (y_n^*, x_n^*)\}$, such a calculation is computationally infeasible. The popular alternative is to use simulation to approximate the distribution. The algorithm is identical to our discussion of Monte Carlo simulation, with the following points of clarification:

- The sample size n used for the simulation is the same as the sample size.
- The random vectors (y_i^*, x_i^*) are drawn randomly from the empirical distribution. This is equivalent to sampling a pair (y_i, x_i) randomly from the sample.

The bootstrap statistic $T_n^* = T_n(y_1^*, x_i^*, ..., y_n^*, x_n^*, F_n)$ is calculated for each bootstrap sample. This is repeated B times. B is known as the number of bootstrap replications. A theory for the determination of the number of bootstrap replications B has been developed by Andrews and Buchinsky (2000). It is desireable for B to be large, so long as the computational costs are reasonable. $B = 1000$ typically suffices.

When the statistic T_n is a function of F, it is typically through dependence on a parameter. For example, the t-ratio $\left(\hat{\theta} - \theta\right) / s(\hat{\theta})$ depends on θ. As the bootstrap statistic replaces F with F_n, it similarly replaces θ with θ_n, the value of θ implied by F_n. Typically $\theta_n = \hat{\theta}$, the parameter estimate. (When in doubt use $\hat{\theta}$.)

at the sample estimate $\hat{\theta}$.

Sampling from the EDF is particularly easy. Since F_n is a discrete probability distribution putting probability mass $1/n$ at each sample point, sampling from the EDF is equivalent to random sampling a pair (y_i, x_i) from the observed data **with replacement**. In consequence, a bootstrap sample $\{y_1^*, x_1^*, ..., y_n^*, x_n^*\}$ will necessarily have some ties and multiple values, which is generally not a problem.

8.4 Bootstrap Estimation of Bias and Variance

The bias of $\hat{\theta}$ is

$$\tau_n = E(\hat{\theta} - \theta_0).$$

Let $T_n(\theta) = \hat{\theta} - \theta$. Then

$$\tau_n = E(T_n(\theta_0)).$$

The bootstrap counterparts are $\hat{\theta}^* = \hat{\theta}(y_1^*, x_1^* ..., y_n^*, x_n^*)$ and $T_n^* = \hat{\theta}^* - \theta_n = \hat{\theta}^* - \hat{\theta}$. The bootstrap estimate of τ_n is

$$\tau_n^* = E(T_n^*).$$

If this is calculated by the simulation described in the previous section, the estimate of τ_n^* is

$$\begin{aligned} \hat{\tau}_n^* &= \frac{1}{B}\sum_{b=1}^{B} T_{nb}^* \\ &= \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b^* - \hat{\theta} \\ &= \overline{\hat{\theta}^*} - \hat{\theta}. \end{aligned}$$

If $\hat{\theta}$ is biased, it might be desirable to construct a biased-corrected estimator (one with reduced bias). Ideally, this would be

$$\tilde{\theta} = \hat{\theta} - \tau_n,$$

but τ_n is unknown. The (estimated) bootstrap biased-corrected estimator is

$$\begin{aligned} \tilde{\theta}^* &= \hat{\theta} - \hat{\tau}_n^* \\ &= \hat{\theta} - (\overline{\hat{\theta}^*} - \hat{\theta}) \\ &= 2\hat{\theta} - \overline{\hat{\theta}^*}. \end{aligned}$$

Note, in particular, that the biased-corrected estimator is *not* $\overline{\hat{\theta}^*}$. Intuitively, the bootstrap makes the following experiment. Suppose that $\hat{\theta}$ is the truth. Then what is the average value of $\hat{\theta}$ calculated from such samples? The answer is $\overline{\hat{\theta}^*}$. If this is lower than $\hat{\theta}$, this suggests that the estimator is *downward-biased*, so a biased-corrected estimator of θ should be *larger* than $\hat{\theta}$, and the best guess is the difference between $\hat{\theta}$ and $\overline{\hat{\theta}^*}$. Similarly if $\overline{\hat{\theta}^*}$ is higher than $\hat{\theta}$, then the estimator is upward-biased and the biased-corrected estimator should be lower than $\hat{\theta}$.

Let $T_n = \hat{\theta}$. The variance of $\hat{\theta}$ is

$$V_n = E(T_n - ET_n)^2.$$

Let $T_n^* = \hat{\theta}^*$. It has variance

$$V_n^* = E(T_n^* - ET_n^*)^2.$$

The simulation estimate is

$$\hat{V}_n^* = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \overline{\hat{\theta}^*}\right)^2.$$

A bootstrap standard error for $\hat{\theta}$ is the square root of the bootstrap estimate of variance, $s(\hat{\beta}) = \sqrt{\hat{V}_n^*}$.

While this standard error may be calculated and reported, it is not clear if it is useful. The primary use of asymptotic standard errors is to construct asymptotic confidence intervals, which are based on the asymptotic normal approximation to the t-ratio. However, the use of the bootstrap presumes that such asymptotic approximations might be poor, in which case the normal approximation is suspected. It appears superior to calculate bootstrap confidence intervals, and we turn to this next.

8.5 Percentile Intervals

For a distribution function $G_n(x, F)$, let $q_n(\alpha, F)$ denote its quantile function. This is the function which solves

$$G_n(q_n(\alpha, F), F) = \alpha.$$

[When $G_n(x, F)$ is discrete, $q_n(\alpha, F)$ may be non-unique, but we will ignore such complications.] Let $q_n(\alpha) = q_n(\alpha, F_0)$ denote the quantile function of the true sampling distribution, and $q_n^*(\alpha) = q_n(\alpha, F_n)$ denote the quantile function of the bootstrap distribution. Note that this function will change depending on the underlying statistic T_n whose distribution is G_n.

Let $T_n = \hat{\theta}$, an estimate of a parameter of interest. In $(1-\alpha)\%$ of samples, $\hat{\theta}$ lies in the region $[q_n(\alpha/2), q_n(1-\alpha/2)]$. This motivates a confidence interval proposed by Efron:

$$C_1 = [q_n^*(\alpha/2), \quad q_n^*(1-\alpha/2)].$$

This is often called the *percentile confidence interval*.

Computationally, the quantile $q_n^*(x)$ is estimated by $\hat{q}_n^*(x)$, the x'th sample quantile of the simulated statistics $\{T_{n1}^*, ..., T_{nB}^*\}$, as discussed in the section on Monte Carlo simulation. The $(1-\alpha)\%$ Efron percentile interval is then $[\hat{q}_n^*(\alpha/2), \quad \hat{q}_n^*(1-\alpha/2)]$.

The interval C_1 is a popular bootstrap confidence interval often used in empirical practice. This is because it is easy to compute, simple to motivate, was popularized by Efron early in the history of the bootstrap, and also has the feature that it is translation invariant. That is, if we define $\phi = f(\theta)$ as the parameter of interest for a monotonic function f, then percentile method applied to this problem will produce the confidence interval $[f(q_n^*(\alpha/2)), \quad f(q_n^*(1-\alpha/2))]$, which is a naturally good property.

However, as we show now, C_1 is in a deep sense very poorly motivated.

It will be useful if we introduce an alternative definition C_1. Let $T_n(\theta) = \hat{\theta} - \theta$ and let $q_n(\alpha)$ be the quantile function of its distribution. (These are the original quantiles, with θ subtracted.) Then C_1 can alternatively be written as

$$C_1 = [\hat{\theta} + q_n^*(\alpha/2), \quad \hat{\theta} + q_n^*(1-\alpha/2)].$$

This is a bootstrap estimate of the "ideal" confidence interval

$$C_1^0 = [\hat{\theta} + q_n(\alpha/2), \quad \hat{\theta} + q_n(1-\alpha/2)].$$

The latter has coverage probability

$$\begin{aligned} P\left(\theta_0 \in C_1^0\right) &= P\left(\hat{\theta} + q_n(\alpha/2) \leq \theta_0 \leq \hat{\theta} + q_n(1-\alpha/2)\right) \\ &= P\left(-q_n(1-\alpha/2) \leq \hat{\theta} - \theta_0 \leq -q_n(\alpha/2)\right) \\ &= G_n(-q_n(\alpha/2), F_0) - G_n(-q_n(1-\alpha/2), F_0) \end{aligned}$$

which generally is not $1-\alpha$! There is one important exception. If $\hat{\theta}-\theta_0$ has a symmetric distribution, then $G_n(-x, F_0) = 1 - G_n(x, F_0)$, so

$$\begin{aligned} P\left(\theta_0 \in C_1^0\right) &= G_n(-q_n(\alpha/2), F_0) - G_n(-q_n(1-\alpha/2), F_0) \\ &= (1 - G_n(q_n(\alpha/2), F_0)) - (1 - G_n(q_n(1-\alpha/2), F_0)) \\ &= \left(1 - \frac{\alpha}{2}\right) - \left(1 - \left(1 - \frac{\alpha}{2}\right)\right) \\ &= 1 - \alpha \end{aligned}$$

and this idealized confidence interval is accurate. Therefore, C_1^0 and C_1 are designed for the case that $\hat{\theta}$ has a symmetric distribution about θ_0.

When $\hat{\theta}$ does not have a symmetric distribution, C_1 may perform quite poorly.

However, by the translation invariance argument presented above, it also follows that if there exists some monotonic transformation $f(\cdot)$ such that $f(\hat{\theta})$ is symmetrically distributed about $f(\theta_0)$, then the idealized percentile bootstrap method will be accurate.

Based on these arguments, many argue that the percentile interval should not be used unless the sampling distribution is close to unbiased and symmetric.

The problems with the percentile method can be circumvented by an alternative method.

Let $T_n(\theta) = \hat{\theta} - \theta$. Then

$$\begin{aligned} 1-\alpha &= P\left(q_n(\alpha/2) \leq T_n(\theta_0) \leq q_n(1-\alpha/2)\right) \\ &= P\left(\hat{\theta} - q_n(1-\alpha/2) \leq \theta_0 \leq \hat{\theta} - q_n(\alpha/2)\right), \end{aligned}$$

so an exact $(1-\alpha)\%$ confidence interval for θ_0 would be

$$C_2^0 = [\hat{\theta} - q_n(1-\alpha/2), \quad \hat{\theta} - q_n(\alpha/2)].$$

This motivates a bootstrap analog

$$C_2 = [\hat{\theta} - q_n^*(1-\alpha/2), \quad \hat{\theta} - q_n^*(\alpha/2)].$$

Notice that generally this is very different from the Efron interval C_1! They coincide in the special case that $G_n^*(x)$ is symmetric about $\hat{\theta}$, but otherwise they differ.

Computationally, this interval can be estimated from a bootstrap simulation by sorting the bootstrap statistics $T_n^* = \left(\hat{\theta}^* - \hat{\theta}\right)$, which are centered at the sample estimate $\hat{\theta}$. These are sorted to yield the quantile estimates $\hat{q}_n^*(.025)$ and $\hat{q}_n^*(.975)$. The 95% confidence interval is then $[\hat{\theta} - \hat{q}_n^*(.975), \quad \hat{\theta} - \hat{q}_n^*(.025)]$.

This confidence interval is discussed in most theoretical treatments of the bootstrap, but is not widely used in practice.

8.6 Percentile-t Equal-Tailed Interval

Suppose we want to test $H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$ at size α. We would set $T_n(\theta) = \left(\hat{\theta} - \theta\right)/s(\hat{\theta})$ and reject H_0 in favor of H_1 if $T_n(\theta_0) < c$, where c would be selected so that

$$P\left(T_n(\theta_0) < c\right) = \alpha.$$

Thus $c = q_n(\alpha)$. Since this is unknown, a bootstrap test replaces $q_n(\alpha)$ with the bootstrap estimate $q_n^*(\alpha)$, and the test rejects if $T_n(\theta_0) < q_n^*(\alpha)$.

Similarly, if the alternative is $H_1 : \theta > \theta_0$, the bootstrap test rejects if $T_n(\theta_0) > q_n^*(1-\alpha)$.

Computationally, these critical values can be estimated from a bootstrap simulation by sorting the bootstrap t-statistics $T_n^* = \left(\hat{\theta}^* - \hat{\theta}\right)/s(\hat{\theta}^*)$. Note, and this is important, that the bootstrap test statistic is centered at the estimate $\hat{\theta}$, and the standard error $s(\hat{\theta}^*)$ is calculated on the bootstrap sample. These t-statistics are sorted to find the estimated quantiles $\hat{q}_n^*(\alpha)$ and/or $\hat{q}_n^*(1-\alpha)$.

Let $T_n(\theta) = \left(\hat{\theta} - \theta\right)/s(\hat{\theta})$. Then

$$\begin{aligned} 1-\alpha &= P\left(q_n(\alpha/2) \le T_n(\theta_0) \le q_n(1-\alpha/2)\right) \\ &= P\left(q_n(\alpha/2) \le \left(\hat{\theta} - \theta_0\right)/s(\hat{\theta}) \le q_n(1-\alpha/2)\right) \\ &= P\left(\hat{\theta} - s(\hat{\theta})q_n(1-\alpha/2) \le \theta_0 \le \hat{\theta} - s(\hat{\theta})q_n(\alpha/2)\right), \end{aligned}$$

so an exact $(1-\alpha)\%$ confidence interval for θ_0 would be

$$C_3^0 = [\hat{\theta} - s(\hat{\theta})q_n(1-\alpha/2), \quad \hat{\theta} - s(\hat{\theta})q_n(\alpha/2)].$$

This motivates a bootstrap analog

$$C_3 = [\hat{\theta} - s(\hat{\theta})q_n^*(1-\alpha/2), \quad \hat{\theta} - s(\hat{\theta})q_n^*(\alpha/2)].$$

This is often called a *percentile-t confidence interval*. It is *equal-tailed* or *central* since the probability that θ_0 is below the left endpoint approximately equals the probability that θ_0 is above the right endpoint, each $\alpha/2$.

Computationally, this is based on the critical values from the one-sided hypothesis tests, discussed above.

8.7 Symmetric Percentile-t Intervals

Suppose we want to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ at size α. We would set $T_n(\theta) = \left(\hat{\theta} - \theta\right)/s(\hat{\theta})$ and reject H_0 in favor of H_1 if $|T_n(\theta_0)| > c$, where c would be selected so that

$$P\left(|T_n(\theta_0)| > c\right) = \alpha.$$

Note that

$$\begin{aligned} P\left(|T_n(\theta_0)| < c\right) &= P\left(-c < T_n(\theta_0) < c\right) \\ &= G_n(c) - G_n(-c) \\ &\equiv \overline{G}_n(c), \end{aligned}$$

which is a symmetric distribution function. The ideal critical value $c = q_n(\alpha)$ solves the equation

$$\overline{G}_n(q_n(\alpha)) = 1 - \alpha.$$

Equivalently, $q_n(\alpha)$ is the $1-\alpha$ quantile of the distribution of $|T_n(\theta_0)|$.

The bootstrap estimate is $q_n^*(\alpha)$, the $1-\alpha$ quantile of the distribution of $|T_n^*|$, or the number which solves the equation

$$\overline{G}_n^*(q_n^*(\alpha)) = G_n^*(q_n^*(\alpha)) - G_n^*(-q_n^*(\alpha)) = 1 - \alpha.$$

Computationally, $q_n^*(\alpha)$ is estimated from a bootstrap simulation by sorting the bootstrap t-statistics $|T_n^*| = \left|\hat{\theta}^* - \hat{\theta}\right|/s(\hat{\theta}^*)$, and taking the upper $\alpha\%$ quantile. The bootstrap test rejects if $|T_n(\theta_0)| > q_n^*(\alpha)$.

Let

$$C_4 = [\hat{\theta} - s(\hat{\theta})q_n^*(\alpha), \quad \hat{\theta} + s(\hat{\theta})q_n^*(\alpha)],$$

where $q_n^*(\alpha)$ is the bootstrap critical value for a two-sided hypothesis test. C_4 is called the symmetric percentile-t interval. It is designed to work well since

$$\begin{aligned} P(\theta_0 \in C_4) &= P\left(\hat{\theta} - s(\hat{\theta})q_n^*(\alpha) \leq \theta_0 \leq \hat{\theta} + s(\hat{\theta})q_n^*(\alpha)\right) \\ &= P\left(|T_n(\theta_0)| < q_n^*(\alpha)\right) \\ &\simeq P\left(|T_n(\theta_0)| < q_n(\alpha)\right) \\ &= 1 - \alpha. \end{aligned}$$

If θ is a vector, then to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ at size α, we would use a Wald statistic

$$W_n(\theta) = n\left(\hat{\theta} - \theta\right)' \hat{V}_\theta^{-1}\left(\hat{\theta} - \theta\right)$$

or some other asymptotically chi-square statistic. Thus here $T_n(\theta) = W_n(\theta)$. The ideal test rejects if $W_n \geq q_n(\alpha)$, where $q_n(\alpha)$ is the $(1-\alpha)\%$ quantile of the distribution of W_n. The bootstrap test rejects if $W_n \geq q_n^*(\alpha)$, where $q_n^*(\alpha)$ is the $(1-\alpha)\%$ quantile of the distribution of

$$W_n^* = n\left(\hat{\theta}^* - \hat{\theta}\right)' \hat{V}_\theta^{*-1}\left(\hat{\theta}^* - \hat{\theta}\right).$$

Computationally, the critical value $q_n^*(\alpha)$ is found as the quantile from simulated values of W_n^*. Note in the simulation that the Wald statistic is a quadratic form in $\left(\hat{\theta}^* - \hat{\theta}\right)$, not $\left(\hat{\theta}^* - \theta_0\right)$. [This is a typical mistake made by practitioners.]

8.8 Asymptotic Expansions

Let T_n be a statistic such that

$$T_n \to_d N(0, v^2). \tag{8.3}$$

If $T_n = \sqrt{n}\left(\hat{\theta} - \theta_0\right)$ then $v = V$ while if T_n is a t-ratio then $v = 1$. Equivalently, writing $T_n \sim G_n(x, F)$ then

$$\lim_{n\to\infty} G_n(x, F) = \Phi\left(\frac{x}{v}\right),$$

or

$$G_n(x, F) = \Phi\left(\frac{x}{v}\right) + o(1). \tag{8.4}$$

While (8.4) says that G_n converges to $\Phi\left(\frac{x}{v}\right)$ as $n \to \infty$, it says nothing, however, about the rate of convergence, or the size of the divergence for any particular sample size n. A better asymptotic approximation may be obtained through an *asymptotic expansion.*

The following notation will be helpful. Let a_n be a sequence.

Definition 8.8.1 *$a_n = o(1)$ if $a_n \to 0$ as $n \to \infty$*

Definition 8.8.2 *$a_n = O(1)$ if $|a_n|$ is uniformly bounded.*

Definition 8.8.3 *$a_n = o(n^{-r})$ if $n^r |a_n| \to 0$ as $n \to \infty$.*

Basically, $a_n = O(n^{-r})$ if it declines to zero like n^{-r}.

We say that a function $g(x)$ is *even* if $g(-x) = g(x)$, and a function $h(x)$ is *odd* if $h(-x) = -h(x)$. The derivative of an even function is odd, and vice-versa.

Theorem 8.8.1 *Under regularity conditions and (8.3),*

$$G_n(x,F) = \Phi\left(\frac{x}{v}\right) + \frac{1}{n^{1/2}}g_1(x,F) + \frac{1}{n}g_2(x,F) + O(n^{-3/2})$$

uniformly over x, where g_1 is an even function of x, and g_2 is an odd function of x. Moreover, g_1 and g_2 are differentiable functions of x and continuous in F relative to the supremum norm on the space of distribution functions.

We can interpret Theorem 8.8.1 as follows. First, $G_n(x,F)$ converges to the normal limit at rate $n^{1/2}$. To a second order of approximation,

$$G_n(x,F) \approx \Phi\left(\frac{x}{v}\right) + n^{-1/2}g_1(x,F).$$

Since the derivative of g_1 is odd, the density function is skewed. To a third order of approximation,

$$G_n(x,F) \approx \Phi\left(\frac{x}{v}\right) + n^{-1/2}g_1(x,F) + n^{-1}g_2(x,F)$$

which adds a symmetric non-normal component to the approximate density (for example, adding leptokurtosis).

8.9 One-Sided Tests

Using the expansion of Theorem 8.8.1, we can assess the accuracy of one-sided hypothesis tests and confidence regions based on an asymptotically normal t-ratio T_n. An asymptotic test is based on $\Phi(x)$.

To the second order, the exact distribution is

$$P\left(T_n < x\right) = G_n(x,F_0) = \Phi(x) + \frac{1}{n^{1/2}}g_1(x,F_0) + O(n^{-1})$$

since $v = 1$. The difference is

$$\begin{aligned}\Phi(x) - G_n(x,F_0) &= \frac{1}{n^{1/2}}g_1(x,F_0) + O(n^{-1}) \\ &= O(n^{-1/2}),\end{aligned}$$

so the order of the error is $O(n^{-1/2})$.

A bootstrap test is based on $G_n^*(x)$, which from Theorem 8.8.1 has the expansion

$$G_n^*(x) = G_n(x,F_n) = \Phi(x) + \frac{1}{n^{1/2}}g_1(x,F_n) + O(n^{-1}).$$

Because $\Phi(x)$ appears in both expansions, the difference between the bootstrap distribution and the true distribution is

$$G_n^*(x) - G_n(x,F_0) = \frac{1}{n^{1/2}}\left(g_1(x,F_n) - g_1(x,F_0)\right) + O(n^{-1}).$$

Since F_n converges to F at rate $\sqrt{n}$, and g_1 is continuous with respect to F, the difference $(g_1(x,F_n) - g_1(x,F_0))$ converges to 0 at rate $\sqrt{n}$. Heuristically,

$$\begin{aligned}g_1(x,F_n) - g_1(x,F_0) &\approx \frac{\partial}{\partial F}g_1(x,F_0)\left(F_n - F_0\right) \\ &= O(n^{-1/2}),\end{aligned}$$

The "derivative" $\frac{\partial}{\partial F} g_1(x, F)$ is only heuristic, as F is a function. We conclude that

$$G_n^*(x) - G_n(x, F_0) = O(n^{-1}),$$

or

$$P\left(T_n^* \leq x\right) = P\left(T_n \leq x\right) + O(n^{-1}),$$

which is an improved rate of convergence over the asymptotic test (which converged at rate $O(n^{-1/2})$). This rate can be used to show that one-tailed bootstrap inference based on the t-ratio achieves a so-called *asymptotic refinement* – the Type I error of the test converges at a faster rate than an analogous asymptotic test.

8.10 Symmetric Two-Sided Tests

If a random variable X has distribution function $H(x) = P(X \leq x)$, then the random variable $|X|$ has distribution function

$$\overline{H}(x) = H(x) - H(-x)$$

since

$$\begin{aligned} P\left(|X| \leq x\right) &= P\left(-x \leq X \leq x\right) \\ &= P\left(X \leq x\right) - P\left(X \leq -x\right) \\ &= H(x) - H(-x). \end{aligned}$$

For example, if $Z \sim N(0,1)$, then $|Z|$ has distribution function

$$\overline{\Phi}(x) = \Phi(x) - \Phi(-x) = 2\Phi(x) - 1.$$

Similarly, if T_n has exact distribution $G_n(x, F)$, then $|T_n|$ has the distribution function

$$\overline{G}_n(x, F) = G_n(x, F) - G_n(-x, F).$$

A two-sided hypothesis test rejects H_0 for large values of $|T_n|$. Since $T_n \to_d Z$, then $|T_n| \to_d |Z| \sim \overline{\Phi}$. Thus asymptotic critical values are taken from the $\overline{\Phi}$ distribution, and exact critical values are taken from the $\overline{G}_n(x, F_0)$ distribution. From Theorem 8.8.1, we can calculate that

$$\begin{aligned} \overline{G}_n(x, F) &= G_n(x, F) - G_n(-x, F) \\ &= \left(\Phi(x) + \frac{1}{n^{1/2}} g_1(x, F) + \frac{1}{n} g_2(x, F)\right) \\ &\quad - \left(\Phi(-x) + \frac{1}{n^{1/2}} g_1(-x, F) + \frac{1}{n} g_2(-x, F)\right) + O(n^{-3/2}) \\ &= \overline{\Phi}(x) + \frac{2}{n} g_2(x, F) + O(n^{-3/2}), \end{aligned} \tag{8.5}$$

where the simplifications are because g_1 is even and g_2 is odd. Hence the difference between the asymptotic distribution and the exact distribution is

$$\overline{\Phi}(x) - \overline{G}_n(x, F_0) = \frac{2}{n} g_2(x, F_0) + O(n^{-3/2}) = O(n^{-1}).$$

The order of the error is $O(n^{-1})$.

Interestingly, the asymptotic two-sided test has a better coverage rate than the asymptotic one-sided test. This is because the first term in the asymptotic expansion, g_1, is an even function, meaning that the errors in the two directions exactly cancel out.

Applying (8.5) to the bootstrap distribution, we find

$$\overline{G}_n^*(x) = \overline{G}_n(x, F_n) = \overline{\Phi}(x) + \frac{2}{n} g_2(x, F_n) + O(n^{-3/2}).$$

Thus the difference between the bootstrap and exact distributions is

$$\begin{aligned} \overline{G}_n^*(x) - \overline{G}_n(x, F_0) &= \frac{2}{n} \left(g_2(x, F_n) - g_2(x, F_0) \right) + O(n^{-3/2}) \\ &= O(n^{-3/2}), \end{aligned}$$

the last equality because F_n converges to F_0 at rate $\sqrt{n}$, and g_2 is continuous in F. Another way of writing this is

$$P\left(|T_n^*| < x\right) = P\left(|T_n| < x\right) + O(n^{-3/2})$$

so the error from using the bootstrap distribution (relative to the true unknown distribution) is $O(n^{-3/2})$. This is in contrast to the use of the asymptotic distribution, whose error is $O(n^{-1})$. Thus a two-sided bootstrap test also achieves an asymptotic refinement, similar to a one-sided test.

A reader might get confused between the two simultaneous effects. Two-sided tests have better rates of convergence than the one-sided tests, and bootstrap tests have better rates of convergence than asymptotic tests.

The analysis shows that there may be a trade-off between one-sided and two-sided tests. Two-sided tests will have more accurate size (Reported Type I error), but one-sided tests might have more power against alternatives of interest. Confidence intervals based on the bootstrap can be asymmetric if based on one-sided tests (equal-tailed intervals) and can therefore be more informative and have smaller length than symmetric intervals. Therefore, the choice between symmetric and equal-tailed confidence intervals is unclear, and needs to be determined on a case-by-case basis.

8.11 Percentile Confidence Intervals

To evaluate the coverage rate of the percentile interval, set $T_n = \sqrt{n}\left(\hat{\theta} - \theta_0\right)$. We know that $T_n \to_d N(0, V)$, which is not pivotal, as it depends on the unknown V. Theorem 8.8.1 shows that a first-order approximation

$$G_n(x, F) = \Phi\left(\frac{x}{v}\right) + O(n^{-1/2}),$$

where $v = \sqrt{V}$, and for the bootstrap

$$G_n^*(x) = G_n(x, F_n) = \Phi\left(\frac{x}{\hat{v}}\right) + O(n^{-1/2}),$$

where $\hat{V} = V(F_n)$ is the bootstrap estimate of V. The difference is

$$\begin{aligned} G_n^*(x) - G_n(x, F_0) &= \Phi\left(\frac{x}{\hat{v}}\right) - \Phi\left(\frac{x}{v}\right) + O(n^{-1/2}) \\ &= -\phi\left(\frac{x}{\hat{v}}\right) \frac{x}{\hat{v}} \left(\hat{v} - v\right) + O(n^{-1/2}) \\ &= O(n^{-1/2}) \end{aligned}$$

Hence the order of the error is $O(n^{-1/2})$.

The good news is that the percentile-type methods (if appropriately used) can yield $\sqrt{n}$-convergent asymptotic inference. Yet these methods do not require the calculation of standard errors! This means that in contexts where standard errors are not available or are difficult to calculate, the percentile bootstrap methods provide an attractive inference method.

The bad news is that the rate of convergence is disappointing. It is no better than the rate obtained from an asymptotic one-sided confidence region. Therefore if standard errors are available, it is unclear if there are any benefits from using the percentile bootstrap over simple asymptotic methods.

Based on these arguments, the theoretical literature (e.g. Hall, 1992, Horowitz, 2002) tends to advocate the use of the percentile-t bootstrap methods rather than percentile methods.

8.12 Bootstrap Methods for Regression Models

The bootstrap methods we have discussed have set $G_n^*(x) = G_n(x, F_n)$, where F_n is the EDF. Any other consistent estimate of F_0 may be used to define a feasible bootstrap estimator. The advantage of the EDF is that it is fully nonparametric, it imposes no conditions, and works in nearly any context. But since it is fully nonparametric, it may be inefficient in contexts where more is known about F. We discuss some bootstrap methods appropriate for the case of a regression model where

$$\begin{aligned} y_i &= x_i'\beta + \varepsilon_i \\ E(e_i \mid x_i) &= 0. \end{aligned}$$

The non-parametric bootstrap distribution resamples the observations (y_i^*, x_i^*) from the EDF, which implies

$$\begin{aligned} y_i^* &= x_i^{*\prime}\hat{\beta} + \varepsilon_i^* \\ E(x_i^*\varepsilon_i^*) &= 0 \end{aligned}$$

but generally

$$E(\varepsilon_i^* \mid x_i^*) \neq 0.$$

The the bootstrap distribution does not impose the regression assumption, and is thus an inefficient estimator of the true distribution (when in fact the regression assumption is true.)

One approach to this problem is to impose the very strong assumption that the error ε_i is independent of the regressor x_i. The advantage is that in this case it is straightforward to construct bootstrap distributions. The disadvantage is that the bootstrap distribution may be a poor approximation when the error is not independent of the regressors.

To impose independence, it is sufficient to sample the x_i^* and ε_i^* independently, and then create $y_i^* = x_i^{*\prime}\hat{\beta} + \varepsilon_i^*$. There are different ways to impose independence. A non-parametric method is to sample the bootstrap errors ε_i^* randomly from the OLS residuals $\{\hat{e}_1, ..., \hat{e}_n\}$. A parametric method is to generate the bootstrap errors ε_i^* from a parametric distribution, such as the normal $\varepsilon_i^* \sim N(0, \hat{\sigma}^2)$.

For the regressors x_i^*, a nonparametric method is to sample the x_i^* randomly from the EDF or sample values $\{x_1, ..., x_n\}$. A parametric method is to sample x_i^* from an estimated parametric distribution. A third approach sets $x_i^* = x_i$. This is equivalent to treating the regressors as *fixed in repeated samples*. If this is done, then all inferential statements are made conditionally on the

observed values of the regressors, which is a valid statistical approach. It does not really matter, however, whether or not the x_i are really "fixed" or random.

The methods discussed above are unattractive for most applications in econometrics because they impose the stringent assumption that x_i and ε_i are independent. Typically what is desirable is to impose only the regression condition $E(\varepsilon_i \mid x_i) = 0$. Unfortunately this is a harder problem.

One proposal which imposes the regression condition without independence is the *Wild Bootstrap*. The idea is to construct a conditional distribution for ε_i^* so that

$$\begin{aligned} E(\varepsilon_i^* \mid x_i) &= 0 \\ E\left(\varepsilon_i^{*2} \mid x_i\right) &= \hat{e}_i^2 \\ E\left(\varepsilon_i^{*3} \mid x_i\right) &= \hat{e}_i^3. \end{aligned}$$

A conditional distribution with these features will preserve the main important features of the data. This can be achieved using a two-point distribution of the form

$$\begin{aligned} P\left(\varepsilon_i^* = \left(\frac{1+\sqrt{5}}{2}\right)\hat{e}_i\right) &= \frac{\sqrt{5}-1}{2\sqrt{5}} \\ P\left(\varepsilon_i^* = \left(\frac{1-\sqrt{5}}{2}\right)\hat{e}_i\right) &= \frac{\sqrt{5}+1}{2\sqrt{5}} \end{aligned}$$

For each x_i, you sample ε_i^* using this two-point distribution.

8.13 Exercises

1. Let $F_n(x)$ denote the EDF of a random sample. Show that

$$\sqrt{n}\left(F_n(x) - F_0(x)\right) \to_d N\left(0, F_0(x)\left(1 - F_0(x)\right)\right).$$

2. Take a random sample $\{y_1, ..., y_n\}$ with $\mu = Ey_i$ and $\sigma^2 = Var(y_i)$. Let the statistic of interest be the sample mean $T_n = \overline{y}_n$. Find the population moments ET_n and $Var(T_n)$. Let $\{y_1^*, ..., y_n^*\}$ be a random sample from the empirical distribution function and let $T_n^* = \overline{y}_n^*$ be its sample mean. Find the bootstrap moments ET_n^* and $Var(T_n^*)$.

3. Consider the following bootstrap procedure for a regression of y_i on x_i. Let $\hat{\beta}$ denote the OLS estimator from the regression of Y on X, and $\hat{e} = Y - X\hat{\beta}$ the OLS residuals.

   (a) Draw a random vector (x^*, e^*) from the pair $\{(x_i, \hat{e}_i) : i = 1, ..., n\}$. That is, draw a random integer i' from $[1, 2, ..., n]$, and set $x^* = x_{i'}$ and $e^* = \hat{e}_{i'}$. Set $y^* = x^{*\prime}\hat{\beta} + e^*$. Draw (with replacement) n such vectors, creating a random bootstrap data set (Y^*, X^*).

   (b) Regress Y^* on X^*, yielding OLS estimates $\hat{\beta}^*$ and any other statistic of interest.

   Show that this bootstrap procedure is (numerically) *identical* to the non-parametric bootstrap.

4. Consider the following bootstrap procedure. Using the non-parametric bootstrap, generate bootstrap samples, calculate the estimate $\hat{\theta}^*$ on these samples and then calculate

$$T_n^* = (\hat{\theta}^* - \hat{\theta})/s(\hat{\theta}),$$

where $s(\hat{\theta})$ is the standard error in the original data. Let $q_n^*(.05)$ and $q_n^*(.95)$ denote the 5% and 95% quantiles of T_n^*, and define the bootstrap confidence interval

$$C = \left[\hat{\theta} - s(\hat{\theta})q_n^*(.95), \quad \hat{\theta} - s(\hat{\theta})q_n^*(.05)\right].$$

Show that C exactly equals the Alternative percentile interval (not the percentile-t interval).

5. You want to test $H_0 : \theta = 0$ against $H_1 : \theta > 0$. The test for H_0 is to reject if $T_n = \hat{\theta}/s(\hat{\theta}) > c$ where c is picked so that Type I error is α. You do this as follows. Using the non-parametric bootstrap, you generate bootstrap samples, calculate the estimates $\hat{\theta}^*$ on these samples and then calculate

$$T_n^* = \hat{\theta}^*/s(\hat{\theta}^*).$$

Let $q_n^*(.95)$ denote the 95% quantile of T_n^*. You replace c with $q_n^*(.95)$, and thus reject H_0 if $T_n = \hat{\theta}/s(\hat{\theta}) > q_n^*(.95)$. What is wrong with this procedure?

6. Suppose that in an application, $\hat{\theta} = 1.2$ and $s(\hat{\theta}) = .2$. Using the non-parametric bootstrap, 1000 samples are generated from the bootstrap distribution, and $\hat{\theta}^*$ is calculated on each sample. The $\hat{\theta}^*$ are sorted, and the 2.5% and 97.5% quantiles of the $\hat{\theta}^*$ are .75 and 1.3, respectively.

   (a) Report the 95% Efron Percentile interval for θ.

   (b) Report the 95% Alternative Percentile interval for θ.

   (c) With the given information, can you report the 95% Percentile-t interval for θ?

7. The datafile *hprice1.dat* contains data on house prices (sales), with variables listed in the file *hprice1.pdf*. Estimate a linear regression of price on the number of bedrooms, lot size, size of house, and the colonial dummy. Calculate 95% confidence intervals for the regression coefficients using both the asymptotic normal approximation and the percentile-t bootstrap.

Chapter 9

Generalized Method of Moments

9.1 Overidentified Linear Model

Consider the linear model

$$\begin{aligned} y_i &= x_i'\beta + e_i \\ &= x_{1i}'\beta_1 + x_{2i}'\beta_2 + e_i \\ E(x_i e_i) &= 0 \end{aligned}$$

where x_{1i} is $k \times 1$ and x_2 is $r \times 1$ with $\ell = k + r$. We know that without further restrictions, an asymptotically efficient estimator of β is the OLS estimator. Now suppose that we are given the information that $\beta_2 = 0$. Now we can write the model as

$$\begin{aligned} y_i &= x_{1i}'\beta_1 + e_i \\ E(x_i e_i) &= 0. \end{aligned}$$

In this case, how should β_1 be estimated? One method is OLS regression of y_i on x_{1i} alone. This method, however, is not necessarily efficient, as their are ℓ restrictions in $E(x_i e_i) = 0$, while β_1 is of dimension $k < \ell$. This situation is called **overidentified**. There are $\ell - k = r$ more moment restrictions than free parameters. We call r the **number of overidentifying restrictions**.

This is a special case of a more general class of moment condition models. Let $g(y, z, x, \beta)$ be an $\ell \times 1$ function of a $k \times 1$ parameter β with $\ell \geq k$ such that

$$Eg(y_i, z_i, x_i, \beta_0) = 0 \tag{9.1}$$

where β_0 is the true value of β. In our previous example, $g(y, x, \beta) = x(y - x_1'\beta)$. In econometrics, this class of models are called **moment condition models**. In the statistics literature, these are known as **estimating equations**.

As an important special case we will devote special attention to linear moment condition models, which can be written as

$$\begin{aligned} y_i &= z_i'\beta + e_i \\ E(x_i e_i) &= 0. \end{aligned}$$

where the dimensions of z_i and x_i are $k \times 1$ and $\ell \times 1$, with $\ell \geq k$. If $k = \ell$ the model is **just identified**, otherwise it is **overidentified**. The variables z_i may be components and functions of x_i, but this is not required. This model falls in the class (9.1) by setting

$$g(y, z, x, \beta_0) = x(y - z'\beta) \tag{9.2}$$

9.2 GMM Estimator

Define the sample analog of (9.2)

$$\overline{g}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\beta) = \frac{1}{n}\sum_{i=1}^{n} x_i\left(y_i - z_i'\beta\right) = \frac{1}{n}\left(X'Y - X'Z\beta\right). \tag{9.3}$$

The method of moments estimator for β is defined as the parameter value which sets $\overline{g}_n(\beta) = 0$, but this is generally not possible when $\ell > k$. The idea of the generalized method of moments (GMM) is to define an estimator which sets $\overline{g}_n(\beta)$ "close" to zero.

For some $\ell \times \ell$ weight matrix $W_n > 0$, let

$$J_n(\beta) = n \cdot \overline{g}_n(\beta)' W_n \overline{g}_n(\beta).$$

This is a non-negative measure of the "length" of the vector $\overline{g}_n(\beta)$. For example, if $W_n = I$, then, $J_n(\beta) = n \cdot \overline{g}_n(\beta)'\overline{g}_n(\beta) = n \cdot |\overline{g}_n(\beta)|^2$, the square of the Euclidean length. The GMM estimator minimizes $J_n(\beta)$.

Definition 9.2.1 $\hat{\beta}_{GMM} = \underset{\beta}{\operatorname{argmin}}\, J_n(\beta)$.

Note that if $k = \ell$, then $\overline{g}_n(\hat{\beta}) = 0$, and the GMM estimator is the MME.

The first order conditions for the GMM estimator are

$$\begin{aligned} 0 &= \frac{\partial}{\partial\beta} J_n(\hat{\beta}) \\ &= 2\frac{\partial}{\partial\beta}\overline{g}_n(\hat{\beta})' W_n \overline{g}_n(\hat{\beta}) \\ &= -2\frac{1}{n}Z'XW_n\frac{1}{n}X'\left(Y - Z\hat{\beta}\right) \end{aligned}$$

so

$$2Z'XW_nX'Z\hat{\beta} = 2Z'XW_nX'Y$$

which establishes the following.

Proposition 9.2.1

$$\hat{\beta}_{GMM} = \left(Z'XW_nX'Z\right)^{-1} Z'XW_nX'Y.$$

While the estimator depends on W_n, the dependence is only up to scale, for if W_n is replaced by cW_n for some $c > 0$, $\hat{\beta}_{GMM}$ does not change.

9.3 Distribution of GMM Estimator

Assume that $W_n \to_p W > 0$. Let

$$Q = E\left(x_i z_i'\right)$$

and

$$\Omega = E\left(x_i x_i' e_i^2\right) = E\left(g_i g_i'\right),$$

where $g_i = x_i e_i$. Then

$$\left(\frac{1}{n}Z'X\right) W_n \left(\frac{1}{n}X'Z\right) \to_p Q'WQ$$

and

$$\left(\frac{1}{n}Z'X\right)W_n\left(\frac{1}{n}X'e\right) \to_d Q'WN(0,\Omega).$$

We conclude:

Theorem 9.3.1 $\sqrt{n}\left(\hat{\beta}-\beta\right) \to_d N(0,V)$, *where*

$$V = \left(Q'WQ\right)^{-1}\left(Q'W\Omega WQ\right)\left(Q'WQ\right)^{-1}.$$

In general, GMM estimators are asymptotically normal with "sandwich form" asymptotic variances.

The optimal weight matrix W_0 is one which minimizes V. This turns out to be $W_0 = \Omega^{-1}$. The proof is left as an exercise. This yields the *efficient GMM* estimator:

$$\hat{\beta} = \left(Z'X\Omega^{-1}X'Z\right)^{-1}Z'X\Omega^{-1}X'Y.$$

Thus we have

Theorem 9.3.2 *For the efficient GMM estimator,* $\sqrt{n}\left(\hat{\beta}-\beta\right) \to_d N\left(0,\left(Q'\Omega^{-1}Q\right)^{-1}\right)$.

This estimator is efficient only in the sense that it is the best (asymptotically) in the class of GMM estimators with this set of moment conditions.

$W_0 = \Omega^{-1}$ is not known in practice, but it can be estimated consistently. For any $W_n \to_p W_0$, we still call $\hat{\beta}$ the efficient GMM estimator, as it has the same asymptotic distribution.

We have described the estimator $\hat{\beta}$ as "efficient GMM" if the optimal (variance minimizing) weight matrix is selected. This is a weak concept of optimality, as we are only considering alternative weight matrices W_n. However, it turns out that the GMM estimator is semiparametrically efficient, as shown by Gary Chamberlain (1987).

If it is known that $E(g_i(\beta)) = 0$, and this is all that is known, this is a semi-parametric problem, as the distribution of the data is unknown. Chamberlain showed that in this context, no semiparametric estimator (one which is consistent globally for the class of models considered) can have a smaller asymptotic variance than $\left(G'\Omega^{-1}G\right)^{-1}$. Since the GMM estimator has this asymptotic variance, it is semiparametrically efficient.

This results shows that in the linear model, no estimator has greater asymptotic efficiency than the efficient linear GMM estimator. No estimator can do better (in this first-order asymptotic sense), without imposing additional assumptions.

9.4 Estimation of the Efficient Weight Matrix

Given any weight matrix $W_n > 0$, the GMM estimator $\hat{\beta}$ is consistent yet inefficient. For example, we can set $W_n = I_\ell$. In the linear model, a better choice is $W_n = (X'X)^{-1}$. Given any such first-step estimator, we can define the residuals $\hat{e}_i = y_i - z_i'\hat{\beta}$ and moment equations $\hat{g}_i = x_i\hat{e}_i = g(w_i,\hat{\beta})$. Construct

$$\overline{g}_n = \overline{g}_n(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\hat{g}_i,$$

$$\hat{g}_i^* = \hat{g}_i - \overline{g}_n,$$

and define

$$W_n = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{g}_i^* \hat{g}_i^{*\prime}\right)^{-1} = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{g}_i \hat{g}_i' - \overline{g}_n \overline{g}_n'\right)^{-1}. \tag{9.4}$$

Then $W_n \to_p \Omega^{-1} = W_0$, and GMM using W_n as the weight matrix is asymptotically efficient.

A common alternative choice is to set

$$W_n = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{g}_i \hat{g}_i'\right)^{-1}$$

which uses the uncentered moment conditions. Since $Eg_i = 0$, these two estimators are asymptotically equivalent under the hypothesis of correct specification. However, Alastair Hall (2000) has shown that the uncentered estimator is a poor choice. When constructing hypothesis tests, under the alternative hypothesis the moment conditions are violated, i.e. $Eg_i \neq 0$, so the uncentered estimator will contain an undesirable bias term and the power of the test will be adversely affected. A simple solution is to use the centered moment conditions to construct the weight matrix, as in (9.4) above.

Here is a simple way to compute the efficient GMM estimator. First, set $W_n = (X'X)^{-1}$, estimate $\hat{\beta}$ using this weight matrix, and construct the residual $\hat{e}_i = y_i - z_i'\hat{\beta}$. Then set $\hat{g}_i = x_i\hat{e}_i$, and let $\hat{g}$ be the associated $n \times \ell$ matrix. Then the efficient GMM estimator is

$$\hat{\beta} = \left(Z'X\left(\hat{g}'\hat{g} - n\overline{g}_n\overline{g}_n'\right)^{-1}X'Z\right)^{-1} Z'X\left(\hat{g}'\hat{g} - n\overline{g}_n\overline{g}_n'\right)^{-1}X'Y.$$

In most cases, when we say "GMM", we actually mean "efficient GMM". There is little point in using an inefficient GMM estimator as it is easy to compute.

An estimator of the asymptotic variance of $\hat{\beta}$ can be seen from the above formula. Set

$$\hat{V} = n\left(Z'X\left(\hat{g}'\hat{g} - n\overline{g}_n\overline{g}_n'\right)^{-1}X'Z\right)^{-1}.$$

Asymptotic standard errors are given by the square roots of the diagonal elements of $\hat{V}$.

There is an important alternative to the two-step GMM estimator just described. Instead, we can let the weight matrix be considered as a function of β. The criterion function is then

$$J(\beta) = n \cdot \overline{g}_n(\beta)'\left(\frac{1}{n}\sum_{i=1}^{n} g_i^*(\beta)g_i^*(\beta)'\right)^{-1}\overline{g}_n(\beta).$$

where

$$g_i^*(\beta) = g_i(\beta) - \overline{g}_n(\beta)$$

The $\hat{\beta}$ which minimizes this function is called the **continuously-updated GMM estimator**, and was introduced by L. Hansen, Heaton and Yaron (1996).

The estimator appears to have some better properties than traditional GMM, but can be numerically tricky to obtain in some cases. This is a current area of research in econometrics.

9.5 GMM: The General Case

In its most general form, GMM applies whenever an economic or statistical model implies the $\ell \times 1$ moment condition

$$E\left(g_i(\beta)\right) = 0.$$

Often, this is *all* that is known. Identification requires $l \geq k = \dim(\beta)$. The GMM estimator minimizes

$$J(\beta) = n \cdot \overline{g}_n(\beta)' W_n \overline{g}_n(\beta)$$

where

$$\overline{g}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\beta)$$

and

$$W_n = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{g}_i \hat{g}_i' - \overline{g}_n \overline{g}_n'\right)^{-1},$$

with $\hat{g}_i = g_i(\tilde{\beta})$ constructed using a preliminary consistent estimator $\tilde{\beta}$, perhaps obtained by first setting $W_n = I$. Since the GMM estimator depends upon the first-stage estimator, often the weight matrix W_n is updated, and then $\hat{\beta}$ recomputed. This estimator can be iterated if needed.

Theorem 9.5.1 *Under general regularity conditions,* $\sqrt{n}\left(\hat{\beta} - \beta\right) \to_d N\left(0, \left(G'\Omega^{-1}G\right)^{-1}\right)$, *where* $\Omega = \left(E\left(g_i g_i'\right)\right)^{-1}$ *and* $G = E\frac{\partial}{\partial \beta'} g_i(\beta)$. *The variance of* $\hat{\beta}$ *may be estimated by* $\left(\hat{G}'\hat{\Omega}^{-1}\hat{G}\right)^{-1}$ *where* $\hat{\Omega} = n^{-1}\sum_i \hat{g}_i^* \hat{g}_i^{*\prime}$ *and* $\hat{G} = n^{-1}\sum_i \frac{\partial}{\partial \beta'} g_i(\hat{\beta})$.

The general theory of GMM estimation and testing was exposited by L. Hansen (1982).

9.6 Over-Identification Test

Overidentified models ($\ell > k$) are special in the sense that there may not be a parameter value β such that the moment condition

$$Eg(w_i, \beta) = 0$$

holds. Thus the model – the overidentifying restrictions – are testable.

For example, take the linear model $y_i = \beta_1' x_{1i} + \beta_2' x_{2i} + e_i$ with $E\left(x_{1i} e_i\right) = 0$ and $E\left(x_{2i} e_i\right) = 0$. It is possible that $\beta_2 = 0$, so that the linear equation may be written as $y_i = \beta_1' x_{1i} + e_i$. However, it is possible that $\beta_2 \neq 0$, and in this case it would be impossible to find a value of β_1 so that both $E\left(x_{1i}\left(y_i - x_{1i}'\beta_1\right)\right) = 0$ and $E\left(x_{2i}\left(y_i - x_{1i}'\beta_1\right)\right) = 0$ hold simultaneously. In this sense an exclusion restriction can be seen as an overidentifying restriction.

Note that $\overline{g}_n \to_p Eg_i$, and thus $\overline{g}_n$ can be used to assess whether or not the hypothesis that $Eg_i = 0$ is true or not. The criterion function at the parameter estimates is

$$\begin{aligned} J &= n\, \overline{g}_n' W_n \overline{g}_n \\ &= n^2 \overline{g}_n' \left(\hat{g}'\hat{g} - n\overline{g}_n \overline{g}_n'\right)^{-1} \overline{g}_n. \end{aligned}$$

is a quadratic form in $\overline{g}_n$, and is thus a natural test statistic for $H_0 : Eg_i = 0$.

Theorem 9.6.1 *(Sargan-Hansen). Under the hypothesis of correct specification, and if the weight matrix is asymptotically efficient,*

$$J = J(\hat{\beta}) \to_d \chi^2_{\ell - k}.$$

The proof of the theorem is left as an exercise. This result was established by Sargan (1958) for a specialized case, and by L. Hansen (1982) for the general case.

The degrees of freedom of the asymptotic distribution are the number of overidentifying restrictions. If the statistic J exceeds the chi-square critical value, we can reject the model. Based on this information alone, it is unclear what is wrong, but it is typically cause for concern. The GMM overidentification test is a very useful by-product of the GMM methodology, and it is advisable to report the statistic J whenever GMM is the estimation method.

When over-identified models are estimated by GMM, it is customary to report the J statistic as a general test of model adequacy.

9.7 Hypothesis Testing: The Distance Statistic

We described before how to construct estimates of the asymptotic covariance matrix of the GMM estimates. These may be used to construct Wald tests of statistical hypotheses.

If the hypothesis is non-linear, a better approach is to directly use the GMM criterion function. This is sometimes called the GMM Distance statistic, and sometimes called a LR-like statistic (the LR is for likelihood-ratio). The idea was first put forward by Newey and West (1987).

For a given weight matrix W_n, the GMM criterion function is

$$J(\beta) = n \cdot \overline{g}_n(\beta)' W_n \overline{g}_n(\beta)$$

For $h : R^k \to R^r$, the hypothesis is

$$H_0 : h(\beta) = 0.$$

The estimates under H_1 are

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, J(\beta)$$

and those under H_0 are

$$\tilde{\beta} = \underset{h(\beta)=0}{\operatorname{argmin}}\, J(\beta).$$

The two minimizing criterion functions are $J(\hat{\beta})$ and $J(\tilde{\beta})$. The GMM distance statistic is the difference

$$D = J(\tilde{\beta}) - J(\hat{\beta}).$$

Proposition 9.7.1 *If the same weight matrix W_n is used for both null and alternative,*

1. $D \geq 0$
2. $D \to_d \chi^2_r$
3. *If h is linear in β, then D equals the Wald statistic.*

If h is non-linear, the Wald statistic can work quite poorly. In contrast, current evidence suggests that the D statistic appears to have quite good sampling properties, and is the preferred test statistic.

Newey and West (1987) suggested to use the same weight matrix W_n for both null and alternative, as this ensures that $D \geq 0$. This reasoning is not compelling, however, and some current research suggests that this restriction is not necessary for good performance of the test.

This test shares the useful feature of LR tests in that it is a natural by-product of the computation of alternative models.

9.8 Conditional Moment Restrictions

In many contexts, the model implies more than an unconditional moment restriction of the form $Eg_i(\beta) = 0$. It implies a conditional moment restriction of the form

$$E\left(e_i(\beta) \mid x_i\right) = 0$$

where $e_i(\beta)$ is some $s \times 1$ function of the observation and the parameters. In many cases, $s = 1$.

It turns out that this conditional moment restriction is much more powerful, and restrictive, than the unconditional moment restriction discussed above.

Our linear model $y_i = z_i'\beta + e_i$ with instruments x_i falls into this class under the stronger assumption $E\left(e_i \mid x_i\right) = 0$. Then $e_i(\beta) = y_i - z_i'\beta$.

It is also helpful to realize that conventional regression models also fall into this class, except that in this case $z_i = x_i$. For example, in linear regression, $e_i(\beta) = y_i - x_i'\beta$, while in a nonlinear regression model $e_i(\beta) = y_i - g(x_i, \beta)$. In a joint model of the conditional mean and variance

$$e_i(\beta, \gamma) = \begin{cases} y_i - x_i'\beta \\ \\ (y_i - x_i'\beta)^2 - f(x_i)'\gamma \end{cases}.$$

Here $s = 2$.

Given a conditional moment restriction, an unconditional moment restriction can always be constructed. That is for any $\ell \times 1$ function $\phi(x_i, \beta)$, we can set $g_i(\beta) = \phi(x_i, \beta)e_i(\beta)$ which satisfies $Eg_i(\beta) = 0$ and hence defines a GMM estimator. The obvious problem is that the class of functions ϕ is infinite. Which should be selected?

This is equivalent to the problem of selection of the best instruments. If x_i is a valid instrument satisfying $E\left(e_i \mid x_i\right) = 0$, then x_i, x_i^2, x_i^3, ..., etc., are all valid instruments. Which should be used?

One solution is to construct an infinite list of potent instruments, and then use the first k instruments. How is k to be determined? This is an area of theory still under development. A recent study of this problem is Donald and Newey (2001).

Another approach is to construct the *optimal instrument*. The form was uncovered by Chamberlain (1987). Take the case $s = 1$. Let

$$R_i = E\left(\frac{\partial}{\partial \beta} e_i(\beta) \mid x_i\right)$$

and

$$\sigma_i^2 = E\left(e_i(\beta)^2 \mid x_i\right).$$

Then the "optimal instrument" is

$$A_i = -\sigma_i^{-2} R_i$$

so the optimal moment is

$$g_i(\beta) = A_i e_i(\beta).$$

Setting $g_i\left(\beta\right)$ to be this choice (which is $k \times 1$, so is just-identified) yields the best GMM estimator possible.

In practice, A_i is unknown, but its form does help us think about construction of optimal instruments.

In the linear model $e_i(\beta) = y_i - z_i'\beta$, note that

$$R_i = -E\left(z_i \mid x_i\right)$$

and

$$\sigma_i^2 = E\left(e_i^2 \mid x_i\right),$$

so

$$A_i = \sigma_i^{-2} E\left(z_i \mid x_i\right).$$

In the case of linear regression, $z_i = x_i$, so $A_i = \sigma_i^{-2} x_i$. Hence efficient GMM is GLS, as we discussed earlier in the course.

In the case of endogenous variables, note that the efficient instrument A_i involves the estimation of the conditional mean of z_i given x_i. In other words, to get the best instrument for z_i, we need the best conditional mean model for z_i given x_i, not just an arbitrary linear projection. The efficient instrument is also inversely proportional to the conditional variance of e_i. This is the same as the GLS estimator; namely that improved efficiency can be obtained if the observations are weighted inversely to the conditional variance of the errors.

9.9 Bootstrap GMM Inference

Let $\hat{\beta}$ be the 2SLS or GMM estimator of β. Using the EDF of (y_i, x_i, z_i), we can apply the bootstrap methods discussed in Chapter 8 to compute estimates of the bias and variance of $\hat{\beta}$, and construct confidence intervals for β, identically as in the regression model. However, caution should be applied when interpreting such results.

A straightforward application of the nonparametric bootstrap works in the sense of consistently achieving the first-order asymptotic distribution. This has been shown by Hahn (1996). However, it fails to achieve an asymptotic refinement when the model is over-identified, jeopardizing the theoretical justification for percentile-t methods. Furthermore, the bootstrap applied J test will yield the wrong answer.

The problem is that in the sample, $\hat{\beta}$ is the "true" value and yet $\overline{g}_n(\hat{\beta}) \neq 0$. Thus according to random variables (y_i^*, x_i^*, z_i^*) drawn from the EDF F_n,

$$E\left(g_i\left(\hat{\beta}\right)\right) = \overline{g}_n(\hat{\beta}) \neq 0.$$

This means that w_i^* do not satisfy the same moment conditions as the population distribution.

A correction suggested by Hall and Horowitz (1996) can solve the problem. Given the bootstrap sample (Y^*, X^*, Z^*), define the bootstrap GMM criterion

$$J^*(\beta) = n \cdot \left(\overline{g}_n^*(\beta) - \overline{g}_n(\hat{\beta})\right)' W_n^* \left(\overline{g}_n^*(\beta) - \overline{g}_n(\hat{\beta})\right)$$

where $\overline{g}_n(\hat{\beta})$ is from the in-sample data, not from the bootstrap data.

Let $\hat{\beta}^*$ minimize $J^*(\beta)$, and define all statistics and tests accordingly. In the linear model, this implies that the bootstrap estimator is

$$\hat{\beta}^* = \left(Z^{*\prime} X^* W_n^* X^{*\prime} Z^*\right)^{-1} \left(Z^{*\prime} X^* W_n^* \left(X^{*\prime} Y^* - X^{\prime} \hat{e}\right)\right).$$

where $\hat{e} = Y - Z\hat{\beta}$ are the in-sample residuals. The bootstrap J statistic is $J^*(\hat{\beta}^*)$.

Brown and Newey (2002) have an alternative solution. They note that we can sample from the observations $\{w_1, ..., w_n\}$ with the empirical likelihood probabilities $\{\hat{p}_i\}$ described in Chapter X. Since $\sum_{i=1}^n \hat{p}_i g_i\left(\hat{\beta}\right) = 0$, this sampling scheme preserves the moment conditions of the model, so no recentering or adjustments are needed. Brown and Newey argue that this bootstrap procedure will be more efficient than the Hall-Horowitz GMM bootstrap.

To date, there are very few empirical applications of bootstrap GMM, as this is a very new area of research.

9.10 Exercises

1. Take the model

$$\begin{aligned} y_i &= x_i'\beta + e_i \\ E\left(x_i e_i\right) &= 0 \\ e_i^2 &= z_i'\gamma + \eta_i \\ E\left(z_i \eta_i\right) &= 0. \end{aligned}$$

Find the method of moments estimators $(\hat{\beta}, \hat{\gamma})$ for (β, γ).

2. Take the single equation

$$\begin{aligned} Y &= Z\beta + e \\ E(e &\mid X) = 0 \end{aligned}$$

Assume $E(e_i^2 \mid x_i) = \sigma^2$. Show that if $\hat{\beta}$ is estimated by GMM with weight matrix $W_n = (X'X)^{-1}$, then

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \to_d N(0, \sigma^2 \left(Q'M^{-1}Q\right)^{-1})$$

where $Q = E(x_i z_i')$ and $M = E(x_i x_i')$.

3. Take the model $y_i = z_i'\beta + e_i$ with $E\left(x_i e_i\right) = 0$. Let $\hat{e}_i = y_i - z_i'\hat{\beta}$ where $\hat{\beta}$ is consistent for β (e.g. a GMM estimator with arbitrary weight matrix). Define the estimate of the optimal GMM weight matrix

$$W_n = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \hat{e}_i^2\right)^{-1}.$$

Show that $W_n \to_p \Omega^{-1}$ where $\Omega = E\left(x_i x_i' e_i^2\right)$.

4. In the linear model estimated by GMM with general weight matrix W, the asymptotic variance of $\hat{\beta}_{GMM}$ is

$$V = \left(Q'WQ\right)^{-1} Q'W\Omega WQ \left(Q'WQ\right)^{-1}$$

   (a) Let V_0 be this matrix when $W = \Omega^{-1}$. Show that $V_0 = \left(Q'\Omega^{-1}Q\right)^{-1}$.

   (b) We want to show that for any W, $V - V_0$ is positive semi-definite (for then V_0 is the smaller possible covariance matrix and $W = \Omega^{-1}$ is the efficient weight matrix). From matrix algebra, we know that $V - V_0$ is positive semi-definite if and only if

$$V_0^{-1} - V^{-1} = A$$

   is positive semi-definite. Write out the matrix A.

   (c) Since Ω is positive definite, there exists a nonsingular matrix C such that $C'C = \Omega^{-1}$. Letting $H = CQ$ and $G = C'^{-1}WQ$, verify that A can be written as

$$A = H'\left(I - G\left(G'G\right)^{-1}G'\right)H.$$

   (d) Show that A is positive semidefinite.
   Hint: The matrix $I - G\left(G'G\right)^{-1}G'$ is symmetric and idempotent, and therefore positive semidefinite.

5. The equation of interest is

$$\begin{aligned} y_i &= g(x_i, \beta) + e_i \\ E(z_i e_i) &= 0. \end{aligned}$$

The observed data is (y_i, x_i, z_i). z_i is $l \times 1$ and β is $k \times 1$, $l \geq k$. Show how to construct an efficient GMM estimator for β.

6. In the linear model $Y = X\beta + e$ with $E(x_i e_i) = 0$, the Generalized Method of Moments (GMM) criterion function for β is defined as

$$J_n(\beta) = \frac{1}{n} (Y - X\beta)' X \hat{\Omega}_n^{-1} X (Y - X\beta) \tag{9.5}$$

where $\hat{e}_i$ are the OLS residuals and $\hat{\Omega}_n = \frac{1}{n} \sum_{i=1}^n x_i x_i' \hat{e}_i^2$. The GMM estimator of β, subject to the restriction $h(\beta) = 0$, is defined as

$$\tilde{\beta} = \underset{h(\beta)=0}{\operatorname{argmin}} J_n(\beta).$$

The GMM test statistic (the distance statistic) of the hypothesis $h(\beta) = 0$ is

$$D = J_n(\tilde{\beta}) = \min_{h(\beta)=0} J_n(\beta). \tag{9.6}$$

(a) Show that you can rewrite $J_n(\beta)$ in (9.5) as

$$J_n(\beta) = \left(\beta - \hat{\beta}\right)' \hat{V}_n^{-1} \left(\beta - \hat{\beta}\right)$$

where

$$\hat{V}_n = \left(X'X\right)^{-1} \left(\sum_{i=1}^n x_i x_i' \hat{e}_i^2\right) \left(X'X\right)^{-1}.$$

(b) Now focus on linear restrictions: $h(\beta) = R'\beta - r$. Thus

$$\tilde{\beta} = \underset{R'\beta - r}{\operatorname{argmin}} J_n(\beta)$$

and hence $R'\tilde{\beta} = r$. Define the Lagrangian $L(\beta, \lambda) = \frac{1}{2} J_n(\beta) + \lambda' (R'\beta - r)$ where λ is $s \times 1$. Show that the minimizer is

$$\begin{aligned} \tilde{\beta} &= \hat{\beta} - \hat{V}_n R \left(R' \hat{V}_n R\right)^{-1} \left(R' \hat{\beta} - r\right) \\ \hat{\lambda} &= \left(R' \hat{V}_n R\right)^{-1} \left(R' \hat{\beta} - r\right). \end{aligned}$$

(c) Show that if $R'\beta = r$ then $\sqrt{n} \left(\tilde{\beta} - \beta\right) \to_d N(0, V_R)$ where

$$V_R = V - VR \left(R'VR\right)^{-1} R'V.$$

(d) Show that in this setting, the distance statistic D in (9.6) equals the Wald statistic.

7. Take the linear model

$$\begin{aligned} y_i &= z_i'\beta + e_i \\ E(x_i e_i) &= 0. \end{aligned}$$

and consider the GMM estimator $\hat{\beta}$ of β. Let

$$J_n = n\overline{g}_n(\hat{\beta})'\hat{\Omega}^{-1}\overline{g}_n(\hat{\beta})$$

denote the test of overidentifying restrictions. Define

$$\begin{aligned} D_n &= I_l - C'\left(\frac{1}{n}X'Z\right)\left(\frac{1}{n}Z'X\hat{\Omega}^{-1}\frac{1}{n}X'Z\right)^{-1}\frac{1}{n}Z'X\hat{\Omega}^{-1}C'^{-1} \\ \overline{g}_n(\beta_0) &= \frac{1}{n}X'e \\ R &= C'E\left(x_i z_i'\right) \end{aligned}$$

Show that $J_n \to_d \chi^2_{l-k}$ as $n \to \infty$ by demonstrating each of the following:

(a) Since $\Omega > 0$, we can write $\Omega^{-1} = CC'$ and $\Omega = C'^{-1}C^{-1}$

(b) $J_n = n\left(C'\overline{g}_n(\hat{\beta})\right)'\left(C'\hat{\Omega}C\right)^{-1}C'\overline{g}_n(\hat{\beta})$

(c) $C'\overline{g}_n(\hat{\beta}) = D_n C'\overline{g}_n(\beta_0)$

(d) $D_n \to_p I_l - R\left(R'R\right)^{-1}R'$

(e) $n^{1/2}C'\overline{g}_n(\beta_0) \to_d N \sim N(0, I_l)$

(f) $J_n \to_d N'\left(I_l - R\left(R'R\right)^{-1}R'\right)N$

(g) $N'\left(I_l - R\left(R'R\right)^{-1}R'\right)N \sim \chi^2_{l-k}$.

Hint: $I_l - R\left(R'R\right)^{-1}R'$ is a projection matrix..

Chapter 10

Empirical Likelihood

10.1 Non-Parametric Likelihood

An alternative to GMM is **empirical likelihood**. The idea is due to Art Owen (1988, 2001) and has been extended to moment condition models by Qin and Lawless (1994). It is a non-parametric analog of likelihood estimation.

The idea is to construct a multinomial distribution $F(p_1, ..., p_n)$ which places probability p_i at each observation. To be a valid multinomial distribution, these probabilities must satisfy the requirements that $p_i \geq 0$ and

$$\sum_{i=1}^{n} p_i = 1. \tag{10.1}$$

Since each observation is observed once in the sample, the log-likelihood function for this multinomial distribution is

$$L_n(p_1, ..., p_n) = \sum_{i=1}^{n} \ln(p_i). \tag{10.2}$$

First let us consider a just-identified model. In this case the moment condition places no additional restrictions on the multinomial distribution. The maximum likelihood estimator of the probabilities $(p_1, ..., p_n)$ are those which maximize the log-likelihood subject to the constraint (10.1). This is equivalent to maximizing

$$\sum_{i=1}^{n} \log(p_i) - \mu \left(\sum_{i=1}^{n} p_i - 1 \right)$$

where μ is a Lagrange multiplier. The n first order conditions are $0 = p_i^{-1} - \mu$. Combined with the constraint (10.1) we find that the MLE is $p_i = n^{-1}$ yielding the log-likelihood $-n \log(n)$.

Now consider the case of an overidentified model with moment condition

$$Eg_i(\beta_0) = 0$$

where g is $\ell \times 1$ and β is $k \times 1$ and for simplicity we write $g_i(\beta) = g(y_i, z_i, x_i, \beta)$. The multinomial distribution which places probability p_i at each observation (y_i, x_i, z_i) will satisfy this condition if and only if

$$\sum_{i=1}^{n} p_i g_i(\beta) = 0 \tag{10.3}$$

The **empirical likelihood estimator** is the value of β which maximizes the multinomial log-likelihood (10.2) subject to the restrictions (10.1) and (10.3).

The Lagrangian for this maximization problem is

$$L_n^*(\beta, p_1, ..., p_n, \lambda, \mu) = \sum_{i=1}^n \ln(p_i) - \mu\left(\sum_{i=1}^n p_i - 1\right) - n\lambda' \sum_{i=1}^n p_i g_i(\beta)$$

where λ and μ are Lagrange multipliers. The first-order-conditions of L_n^* with respect to p_i, μ, and λ are

$$\begin{aligned} \frac{1}{p_i} &= \mu + n\lambda' g_i(\beta) \\ \sum_{i=1}^n p_i &= 1 \\ \sum_{i=1}^n p_i g_i(\beta) &= 0. \end{aligned}$$

Multiplying the first equation by p_i, summing over i, and using the second and third equations, we find $\mu = n$ and

$$p_i = \frac{1}{n\left(1 + \lambda' g_i(\beta)\right)}.$$

Substituting into L_n^* we find

$$R_n(\beta, \lambda) = -n\ln(n) - \sum_{i=1}^n \ln\left(1 + \lambda' g_i(\beta)\right). \tag{10.4}$$

For given β, the Lagrange multiplier $\lambda(\beta)$ minimizes $R_n(\beta, \lambda)$:

$$\lambda(\beta) = \underset{\lambda}{\operatorname{argmin}}\, R_n(\beta, \lambda). \tag{10.5}$$

This minimization problem is the dual of the constrained maximization problem. The solution (when it exists) is well defined since $R_n(\beta, \lambda)$ is a convex function of λ. The solution cannot be obtained explicitly, but must be obtained numerically (see section 6.5). This yields the (profile) empirical log-likelihood function for β.

$$\begin{aligned} L_n(\beta) &= R_n(\beta, \lambda(\beta)) \\ &= -n\ln(n) - \sum_{i=1}^n \ln\left(1 + \lambda(\beta)' g_i(\beta)\right) \end{aligned}$$

The EL estimate $\hat{\beta}$ is the value which maximizes $L_n(\beta)$, or equivalently minimizes its negative

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\left[-L_n(\beta)\right] \tag{10.6}$$

Numerical methods are required for calculation of $\hat{\beta}$. (see section 6.5)

As a by-product of estimation, we also obtain the Lagrange multiplier $\hat{\lambda} = \lambda(\hat{\beta})$, probabilities

$$\hat{p}_i = \frac{1}{n\left(1 + \hat{\lambda}' g_i\left(\hat{\beta}\right)\right)}.$$

and maximized empirical likelihood

$$\hat{L}_n = \sum_{i=1}^n \ln(\hat{p}_i). \tag{10.7}$$

10.2 Asymptotic Distribution of EL Estimator

Define

$$G_i(\beta) = \frac{\partial}{\partial \beta'} g_i(\beta) \tag{10.8}$$

$$G = EG_i(\beta_0)$$

$$\Omega = E\left(g_i(\beta_0) g_i(\beta_0)'\right)$$

and

$$V = \left(G'\Omega^{-1}G\right)^{-1} \tag{10.9}$$

$$V_\lambda = \Omega - G\left(G'\Omega^{-1}G\right)^{-1} G' \tag{10.10}$$

For example, in the linear model, $G_i(,\beta) = -x_i z_i'$, $G = -E(x_i z_i')$, and $\Omega = E\left(x_i x_i' e_i^2\right)$.

Theorem 10.2.1 *Under regularity conditions,*

$$\begin{aligned} \sqrt{n}\left(\hat{\beta} - \beta_0\right) &\to^d \quad N(0, V) \\ \sqrt{n}\hat{\lambda} &\to^d \quad \Omega^{-1} N(0, V_\lambda) \end{aligned}$$

where V and V_λ are defined in (10.9) and (10.10), and $\sqrt{n}\left(\hat{\beta} - \beta_0\right)$ and $\sqrt{n}\hat{\lambda}$ are asymptotically independent.

The asymptotic variance V for $\hat{\beta}$ is the same as for efficient GMM. Thus the EL estimator is asymptotically efficient.

Proof. $(\hat{\beta}, \hat{\lambda})$ jointly solve

$$0 = \frac{\partial}{\partial \lambda} R_n(\beta, \lambda) = -\sum_{i=1}^n \frac{g_i\left(\hat{\beta}\right)}{\left(1 + \hat{\lambda}' g_i\left(\hat{\beta}\right)\right)} \tag{10.11}$$

$$0 = \frac{\partial}{\partial \beta} R_n(\beta, \lambda) = -\sum_{i=1}^n \frac{G_i\left(,\hat{\beta}\right)' \lambda}{1 + \hat{\lambda}' g_i\left(\hat{\beta}\right)}. \tag{10.12}$$

Let $G_n = \frac{1}{n}\sum_{i=1}^n G_i(\beta_0)$, $\overline{g}_n = \frac{1}{n}\sum_{i=1}^n g_i(\beta_0)$ and $\Omega_n = \frac{1}{n}\sum_{i=1i}^n g(\beta_0) g_i(\beta_0)'$.

Expanding (10.12) around $\beta = \beta_0$ and $\lambda = \lambda_0 = 0$ yields

$$0 \simeq G_n'\left(\hat{\lambda} - \lambda_0\right). \tag{10.13}$$

Expanding (10.11) around $\beta = \beta_0$ and $\lambda = \lambda_0 = 0$ yields

$$0 \simeq -\overline{g}_n - G_n\left(\hat{\beta} - \beta_0\right) + \Omega_n \hat{\lambda} \tag{10.14}$$

Premultiplying by $G_n'\Omega_n^{-1}$ and using (10.13) yields

$$\begin{aligned} 0 &\simeq -G_n'\Omega_n^{-1}\overline{g}_n - G_n'\Omega_n^{-1}G_n\left(\hat{\beta} - \beta_0\right) + G_n'\Omega_n^{-1}\Omega_n\hat{\lambda} \\ &= -G_n'\Omega_n^{-1}\overline{g}_n - G_n'\Omega_n^{-1}G_n\left(\hat{\beta} - \beta_0\right) \end{aligned}$$

Solving for $\hat{\beta}$ and using the WLLN and CLT yields

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}-\beta_0\right) &\simeq -\left(G_n'\Omega_n^{-1}G_n\right)^{-1}G_n'\Omega_n^{-1}\sqrt{n}\overline{g}_n \\ &\to^d \left(G'\Omega^{-1}G\right)^{-1}G'\Omega^{-1}N\left(0,\Omega\right) \\ &= N\left(0,V\right)\end{aligned} \tag{10.15}$$

Solving (10.14) for $\hat{\lambda}$ and using (10.15) yields

$$\begin{aligned}\sqrt{n}\hat{\lambda} &\simeq \Omega_n^{-1}\left(I-G_n\left(G_n'\Omega_n^{-1}G_n\right)^{-1}G_n'\Omega_n^{-1}\right)\sqrt{n}\overline{g}_n \\ \to^d &\quad \Omega^{-1}\left(I-G\left(G'\Omega^{-1}G\right)^{-1}G'\Omega^{-1}\right)N\left(0,\Omega\right) \\ &= \Omega^{-1}N\left(0,V_\lambda\right)\end{aligned} \tag{10.16}$$

Furthermore, since

$$G'\left(I-\Omega^{-1}G\left(G'\Omega^{-1}G\right)^{-1}G'\right)=0$$

$\sqrt{n}\left(\hat{\beta}-\beta_0\right)$ and $\sqrt{n}\hat{\lambda}$ are asymptotically uncorrelated and hence independent. ■

Chamberlain (1987) showed that V is the semiparametric efficiency bound for β in the overidentified moment condition model. This means that no consistent estimator for this class of models can have a lower asymptotic variance than V. Since the EL estimator achieves this bound, it is an asymptotically efficient estimator for β.

10.3 Overidentifying Restrictions

In a parametric likelihood context, tests are based on the difference in the log likelihood functions. The same statistic can be constructed for empirical likelihood. Twice the difference between the unrestricted empirical likelihood $-n\log\left(n\right)$ and the maximized empirical likelihood for the model (10.7) is

$$LR_n=\sum_{i=1}^{n}2\ln\left(1+\hat{\lambda}'g_i\left(\hat{\beta}\right)\right). \tag{10.17}$$

Theorem 10.3.1 *If* $Eg(w_i,\beta_0)=0$ *then* $LR_n\to_d\chi^2_{\ell-k}$.

The EL overidentification test is similar to the GMM overidentification test. They are asymptotically first-order equivalent, and have the same interpretation. The overidentification test is a very useful by-product of EL estimation, and it is advisable to report the statistic LR_n whenever EL is the estimation method.

Proof. First, by a Taylor expansion, (10.15), and (10.16),

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g\left(w_i,\hat{\beta}\right) &\simeq \sqrt{n}\left(\overline{g}_n+G_n\left(\hat{\beta}-\beta_0\right)\right) \\ &\simeq \left(I-G_n\left(G_n'\Omega_n^{-1}G_n\right)^{-1}G_n'\Omega_n^{-1}\right)\sqrt{n}\overline{g}_n \\ &\simeq \Omega_n\sqrt{n}\hat{\lambda}.\end{aligned}$$

Second, since $\ln(1+x) \simeq x - x^2/2$ for x small,

$$\begin{aligned} LR_n &= \sum_{i=1}^{n} 2\ln\left(1+\hat{\lambda}' g_i\left(\hat{\beta}\right)\right) \\ &\simeq 2\hat{\lambda}' \sum_{i=1}^{n} g_i\left(\hat{\beta}\right) - \hat{\lambda}' \sum_{i=1}^{n} g_i\left(\hat{\beta}\right) g_i\left(\hat{\beta}\right)' \hat{\lambda} \\ &\simeq n\hat{\lambda}' \Omega_n \hat{\lambda} \\ \to_d & \quad N(0, V_\lambda)' \, \Omega^{-1} N(0, V_\lambda) \\ &= \chi^2_{\ell-k} \end{aligned}$$

where the proof of the final equality is left as an exercise. ■

10.4 Testing

Let the maintained model be

$$Eg_i(\beta) = 0 \tag{10.18}$$

where g is $\ell \times 1$ and β is $k \times 1$. By "maintained" we mean that the overidentfying restrictions contained in (10.18) are assumed to hold and are not being challenged (at least for the test discussed in this section). The hypothesis of interest is

$$h(\beta) = 0.$$

where $h : R^k \to R^a$. The restricted EL estimator and likelihood are the values which solve

$$\begin{aligned} \tilde{\beta} &= \underset{h(\beta)=0}{\operatorname{argmax}} L_n(\beta) \\ \tilde{L}_n &= L_n(\tilde{\beta}) = \max_{h(\beta)=0} L_n(\beta). \end{aligned}$$

Fundamentally, the restricted EL estimator $\tilde{\beta}$ is simply an EL estimator with $\ell - k + a$ overidentifying restrictions, so there is no fundamental change in the distribution theory for $\tilde{\beta}$ relative to $\hat{\beta}$. To test the hypothesis $h(\beta)$ while maintaining (10.18), the simple overidentifying restrictions test (10.17) is not appropriate. Instead we use the difference in log-likelihoods:

$$LR_n = 2\left(\hat{L}_n - \tilde{L}_n\right).$$

This test statistic is a natural analog of the GMM distance statistic.

Theorem 10.4.1 *Under (10.18) and $H_0 : h(\beta) = 0$, $LR_n \to_d \chi^2_a$.*

The proof of this result is more challenging and is omitted.

10.5 Numerical Computation

Gauss code which implements the methods discussed below can be found at

http://www.ssc.wisc.edu/~bhansen/progs/elike.prc

Derivatives

The numerical calculations depend on derivatives of the dual likelihood function (10.4). Define

$$\begin{aligned} g_i^*(\beta,\lambda) &= \frac{g_i(\beta)}{\left(1+\lambda' g_i(\beta)\right)} \\ G_i^*(\beta,\lambda) &= \frac{G_i(\beta)'\lambda}{1+\lambda' g_i(\beta)} \end{aligned}$$

The first derivatives of (10.4) are

$$\begin{aligned} R_\lambda &= \frac{\partial}{\partial\lambda} R_n(\beta,\lambda) = -\sum_{i=1}^n g_i^*(\beta,\lambda) \\ R_\beta &= \frac{\partial}{\partial\beta} R_n(\beta,\lambda) = -\sum_{i=1}^n G_i^*(\beta,\lambda). \end{aligned}$$

The second derivatives are

$$\begin{aligned} R_{\lambda\lambda} &= \frac{\partial^2}{\partial\lambda\partial\lambda'} R_n(\beta,\lambda) = \sum_{i=1}^n g_i^*(\beta,\lambda)\, g_i^*(\beta,\lambda)' \\ R_{\lambda\beta} &= \frac{\partial^2}{\partial\lambda\partial\beta'} R_n(\beta,\lambda) = \sum_{i=1}^n \left(g_i^*(\beta,\lambda)\, G_i^*(\beta,\lambda)' - \frac{G_i(\beta)}{1+\lambda' g_i(\beta)} \right) \\ R_{\beta\beta} &= \frac{\partial^2}{\partial\beta\partial\beta'} R_n(\beta,\lambda) = \sum_{i=1}^n \left(G_i^*(\beta,\lambda)\, G_i^*(\beta,\lambda)' - \frac{\frac{\partial^2}{\partial\beta\partial\beta'}\left(g_i(\beta)'\lambda\right)}{1+\lambda' g_i(\beta)} \right) \end{aligned}$$

Inner Loop

The so-called "inner loop" solves (10.5) for given β. The modified Newton method takes a quadratic approximation to $R_n(\beta,\lambda)$ yielding the iteration rule

$$\lambda_{j+1} = \lambda_j - \delta \left(R_{\lambda\lambda}(\beta,\lambda_j)\right)^{-1} R_\lambda(\beta,\lambda_j). \tag{10.19}$$

where $\delta > 0$ is a scalar steplength (to be discussed next). The starting value λ_1 can be set to the zero vector. The iteration (10.19) is continued until the gradient $R_\lambda(\beta,\lambda_j)$ is smaller than some prespecified tolerance.

Efficient convergence requires a good choice of steplength δ. One method uses the following quadratic approximation. Set $\delta_0 = 0$, $\delta_1 = \frac{1}{2}$ and $\delta_2 = 1$. For $p = 0, 1, 2$, set

$$\begin{aligned} \lambda_p &= \lambda_j - \delta_p \left(R_{\lambda\lambda}(\beta,\lambda_j)\right)^{-1} R_\lambda(\beta,\lambda_j)) \\ R_p &= R_n(\beta,\lambda_p) \end{aligned}$$

A quadratic function can be fit exactly through these three points. The value of δ which minimizes this quadratic is

$$\hat{\delta} = \frac{R_2 + 3R_0 - 4R_1}{4R_2 + 4R_0 - 8R_1}.$$

yielding the steplength to be plugged into (10.19)..

A complication is that λ must be constrained so that $0 \le p_i \le 1$ which holds if

$$n\left(1+\lambda' g_i(\beta)\right) \ge 1 \tag{10.20}$$

for all i. If (10.20) fails, the stepsize δ needs to be decreased.

Outer Loop

The outer loop is the minimization (10.6). This can be done by the modified Newton method described in the previous section. The gradient for (10.6) is

$$L_\beta = \frac{\partial}{\partial \beta} L_n(\beta) = \frac{\partial}{\partial \beta} R_n(\beta, \lambda) = R_\beta + \lambda_\beta' R_\lambda = R_\beta$$

since $R_\lambda(\beta, \lambda) = 0$ at $\lambda = \lambda(\beta)$, where

$$\lambda_\beta = \frac{\partial}{\partial \beta'} \lambda(\beta) = -R_{\lambda\lambda}^{-1} R_{\lambda\beta},$$

the second equality following from the implicit function theorem applied to $R_\lambda(\beta, \lambda(\beta)) = 0$.

The Hessian for (10.6) is

$$\begin{aligned} L_{\beta\beta} &= -\frac{\partial}{\partial \beta \partial \beta'} L_n(\beta) \\ &= -\frac{\partial}{\partial \beta'} \left[R_\beta(\beta, \lambda(\beta)) + \lambda_\beta' R_\lambda(\beta, \lambda(\beta)) \right] \\ &= -\left(R_{\beta\beta}(\beta, \lambda(\beta)) + R_{\lambda\beta}' \lambda_\beta + \lambda_\beta' R_{\lambda\beta} + \lambda_\beta' R_{\lambda\lambda} \lambda_\beta \right) \\ &= R_{\lambda\beta}' R_{\lambda\lambda}^{-1} R_{\lambda\beta} - R_{\beta\beta}. \end{aligned}$$

It is not guaranteed that $L_{\beta\beta} > 0$. If not, the eigenvalues of $L_{\beta\beta}$ should be adjusted so that all are positive. The Newton iteration rule is

$$\beta_{j+1} = \beta_j - \delta L_{\beta\beta}^{-1} L_\beta$$

where δ is a scalar stepsize, and the rule is iterated until convergence.

Chapter 11

Endogeneity

We say that there is endogeneity in the linear model $y = z_i'\beta + e_i$ if β is the parameter of interest and $E(z_i e_i) \neq 0$. This cannot happen if β is defined by linear projection, so requires a structural interpretation. The coefficient β must have meaning separately from the definition of a conditional mean or linear projection.

Example: Measurement error in the regressor. Suppose that (y_i, x_i^*) are joint random variables, $E(y_i \mid x_i^*) = x_i^{*\prime}\beta$ is linear, β is the parameter of interest, and x_i^* is not observed. Instead we observe $x_i = x_i^* + u_i$ where u_i is an $k \times 1$ measurement error, independent of y_i and x_i^*. Then

$$\begin{aligned} y_i &= x_i^{*\prime}\beta + e_i \\ &= (x_i - u_i)'\beta + e_i \\ &= x_i'\beta + v_i \end{aligned}$$

where

$$v_i = e_i - u_i'\beta.$$

The problem is that

$$E(x_i v_i) = E\left[(x_i^* + u_i)\left(e_i - u_i'\beta\right)\right] = -E\left(u_i u_i'\right)\beta \neq 0$$

if $\beta \neq 0$ and $E(u_i u_i') \neq 0$. It follows that if $\hat{\beta}$ is the OLS estimator, then

$$\hat{\beta} \to_p \beta^* = \beta - \left(E\left(x_i x_i'\right)\right)^{-1} E\left(u_i u_i'\right)\beta \neq \beta.$$

This is called **measurement error bias**.

Example: Supply and Demand. The variables q_i and p_i (quantity and price) are determined jointly by the demand equation

$$q_i = -\beta_1 p_i + e_{1i}$$

and the supply equation

$$q_i = \beta_2 p_i + e_{2i}.$$

Assume that $e_i = \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix}$ is iid, $Ee_i = 0$, $\beta_1 + \beta_2 = 1$ and $Ee_i e_i' = I_2$ (the latter for simplicity). The question is, if we regress q_i on p_i, what happens?

It is helpful to solve for q_i and p_i in terms of the errors. In matrix notation,

$$\begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix} \begin{pmatrix} q_i \\ p_i \end{pmatrix} = \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix}$$

so

$$\begin{aligned}\begin{pmatrix} q_i \\ p_i \end{pmatrix} &= \begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix}^{-1} \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \\ &= \begin{bmatrix} \beta_2 & \beta_1 \\ 1 & -1 \end{bmatrix} \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \\ &= \begin{pmatrix} \beta_2 e_{1i} + \beta_1 e_{2i} \\ (e_{1i} - e_{2i}) \end{pmatrix}.\end{aligned}$$

The projection of q_i on p_i yields

$$\begin{aligned} q_i &= \beta^* p_i + \varepsilon_i \\ E(p_i \varepsilon_i) &= 0 \end{aligned}$$

where

$$\beta^* = \frac{E(p_i q_i)}{E(p_i^2)} = \frac{\beta_2 - \beta_1}{2}$$

Hence if it is estimated by OLS, $\hat{\beta} \to_p \beta^*$, which does not equal either β_1 or β_2. This is called **simultaneous equations bias**.

11.1 Instrumental Variables

Let the equation of interest be

$$y_i = z_i' \beta + e_i \tag{11.1}$$

where z_i is $k \times 1$, and assume that $E(z_i e_i) \neq 0$ so there is **endogeneity**. We call (11.1) the structural equation. In matrix notation, this can be written as

$$Y = Z\beta + e. \tag{11.2}$$

Any solution to the problem of endogeneity requires additional information which we call **instruments**.

Definition 11.1.1 *The $\ell \times 1$ random vector x_i is an instrumental variable for (11.1) if $E(x_i e_i) = 0$.*

In a typical set-up, some regressors in z_i will be uncorrelated with e_i (for example, at least the intercept). Thus we make the partition

$$z_i = \begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix} \begin{matrix} k_1 \\ k_2 \end{matrix} \tag{11.3}$$

where $E(z_{1i} e_i) = 0$ yet $E(z_{2i} e_i) \neq 0$. We call z_{1i} exogenous and z_{2i} endogenous. By the above definition, z_{1i} is an instrumental variable for (11.1), so should be included in x_i. So we have the partition

$$x_i = \begin{pmatrix} z_{1i} \\ x_{2i} \end{pmatrix} \begin{matrix} k_1 \\ \ell_2 \end{matrix} \tag{11.4}$$

where $z_{1i} = x_{1i}$ are the **included exogenous variables**, and x_{2i} are the **excluded exogenous variables**. That is x_{2i} are variables which could be included in the equation for y_i (in the sense

that they are uncorrelated with e_i) yet can be *excluded*, as they would have true zero coefficients in the equation.

The model is **just-identified** if $\ell = k$ (i.e., if $\ell_2 = k_2$) and **over-identified** if $\ell > k$ (i.e., if $\ell_2 > k_2$).

We have noted that any solution to the problem of endogeneity requires instruments. This does not mean that valid instruments actually exist.

11.2 Reduced Form

The reduced form relationship between the variables or "regressors" z_i and the instruments x_i is found by linear projection. Let

$$\Gamma = E\left(x_i x_i'\right)^{-1} E\left(x_i z_i'\right)$$

be the $\ell \times k$ matrix of coefficients from a projection of z_i on x_i, and define

$$u_i = z_i - x_i'\Gamma$$

as the projection error. Then the reduced form linear relationship between z_i and x_i is

$$z_i = \Gamma' x_i + u_i. \tag{11.5}$$

In matrix notation, we can write (11.5) as

$$Z = X\Gamma + u \tag{11.6}$$

where u is $n \times k$.

By construction,

$$E(x_i u_i') = 0,$$

so (11.5) is a projection and can be estimated by OLS:

$$\begin{aligned} Z &= X\hat{\Gamma} + \hat{u} \\ \hat{\Gamma} &= \left(X'X\right)^{-1}\left(X'Z\right). \end{aligned}$$

Substituting (11.6) into (11.2), we find

$$\begin{aligned} Y &= (X\Gamma + u)\beta + e \\ &= X\lambda + v, \end{aligned} \tag{11.7}$$

where

$$\lambda = \Gamma\beta \tag{11.8}$$

and

$$v = u\beta + e.$$

Observe that

$$E\left(x_i v_i\right) = E\left(x_i u_i'\right)\beta + E\left(x_i e_i\right) = 0.$$

Thus (11.7) is a projection equation and may be estimated by OLS. This is

$$\begin{aligned} Y &= X\hat{\lambda} + \hat{v}, \\ \hat{\lambda} &= \left(X'X\right)^{-1}\left(X'Y\right) \end{aligned}$$

The equation (11.7) is the reduced form for Y. (11.6) and (11.7) together are the **reduced form equations** for the system

$$\begin{aligned} Y &= X\lambda + v \\ Z &= X\Gamma + u. \end{aligned}$$

As we showed above, OLS yields the reduced-form estimates $(\hat{\lambda}, \hat{\Gamma})$

11.3 Identification

The structural parameter β relates to (λ, Γ) through (11.8). The parameter β is **identified**, meaning that it can be recovered from the reduced form, if

$$rank(\Gamma) = k. \tag{11.9}$$

Assume that (11.9) holds. If $\ell = k$, then $\beta = \Gamma^{-1}\lambda$. If $\ell > k$, then for any $W > 0$, $\beta = (\Gamma' W \Gamma)^{-1} \Gamma' W \lambda$.

If (11.9) is not satisfied, then β cannot be recovered from (λ, Γ). Note that a necessary (although not sufficient) condition for (11.9) is $\ell \geq k$.

Since X and Z have the common variables X_1, we can rewrite some of the expressions. Using (11.3) and (11.4) to make the matrix partitions $X = [X_1, X_2]$ and $Z = [X_1, Z_2]$, we can partition Γ as

$$\begin{aligned} \Gamma &= \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} \\ &= \begin{bmatrix} I & \Gamma_{12} \\ 0 & \Gamma_{22} \end{bmatrix} \end{aligned}$$

(11.6) can be rewritten as

$$\begin{aligned} Z_1 &= X_1 \\ Z_2 &= X_1\Gamma_{12} + X_2\Gamma_{22} + u_2. \end{aligned} \tag{11.10}$$

β is identified if $rank(\Gamma) = k$, which is true if and only if $rank(\Gamma_{22}) = k_2$ (by the upper-diagonal structure of Γ). Thus the key to identification of the model rests on the $\ell_2 \times k_2$ matrix Γ_{22} in (11.10).

11.4 Estimation

The model can be written as

$$\begin{aligned} y_i &= z_i'\beta + e_i \\ E(x_i e_i) &= 0 \end{aligned}$$

or

$$\begin{aligned} Eg(w_i, \beta) &= 0 \\ g(w_i, \beta) &= x_i \left(y_i - z_i'\beta\right). \end{aligned}$$

This a moment condition model. Appropriate estimators include GMM and EL. The estimators and distribution theory developed in those Chapter 8 and 9 directly apply. Recall that the GMM estimator, for given weight matrix W_n, is

$$\hat{\beta} = \left(Z'XW_nX'Z\right)^{-1} Z'XW_nX'Y.$$

11.5 Special Cases: IV and 2SLS

If the model is just-identified, so that $k = \ell$, then the formula for GMM simplifies. We find that

$$\begin{aligned}\hat{\beta} &= \left(Z'XW_nX'Z\right)^{-1} Z'XW_nX'Y \\ &= \left(X'Z\right)^{-1} W_n^{-1} \left(Z'X\right)^{-1} Z'XW_nX'Y \\ &= \left(X'Z\right)^{-1} X'Y\end{aligned}$$

This estimator is often called the **instrumental variables estimator** (IV) of β, where X is used as an instrument for Z. Observe that the weight matrix W_n has disappeared. In the just-identified case, the weight matrix places no role. This is also the MME estimator of β, and the EL estimator. Another interpretation stems from the fact that since $\beta = \Gamma^{-1}\lambda$, we can construct the **Indirect Least Squares** (ILS) estimator:

$$\begin{aligned}\hat{\beta} &= \hat{\Gamma}^{-1}\hat{\lambda} \\ &= \left(\left(X'X\right)^{-1}\left(X'Z\right)\right)^{-1}\left(\left(X'X\right)^{-1}\left(X'Y\right)\right) \\ &= \left(X'Z\right)^{-1}\left(X'X\right)\left(X'X\right)^{-1}\left(X'Y\right) \\ &= \left(X'Z\right)^{-1}\left(X'Y\right).\end{aligned}$$

which again is the IV estimator.

Recall that the optimal weight matrix is an estimate of the inverse of $\Omega = E\left(x_i x_i' e_i^2\right)$. In the special case that $E\left(e_i^2 \mid x_i\right) = \sigma^2$ (homoskedasticity), then $\Omega = E\left(x_i x_i'\right)\sigma^2 \propto E\left(x_i x_i'\right)$ suggesting the weight matrix $W_n = \left(X'X\right)^{-1}$. Using this choice, the GMM estimator equals

$$\hat{\beta}_{2SLS} = \left(Z'X\left(X'X\right)^{-1}X'Z\right)^{-1} Z'X\left(X'X\right)^{-1}X'Y$$

This is called the **two-stage-least squares** (2SLS) estimator. It was originally proposed by Theil (1953) and Basmann (1957), and is the classic estimator for linear equations with instruments. Under the homoskedasticity assumption, the 2SLS estimator is efficient GMM, but otherwise it is inefficient.

It is useful to observe that writing

$$\begin{aligned}P_X &= X\left(X'X\right)^{-1}X', \\ \hat{Z} &= P_X Z = X\left(X'X\right)^{-1}X'Z,\end{aligned}$$

then

$$\begin{aligned}\hat{\beta} &= \left(Z'P_X Z\right)^{-1} Z'P_X Y \\ &= \left(\hat{Z}'\hat{Z}\right)\hat{Z}'Y.\end{aligned}$$

The source of the "two-stage" name is since it can be computed as follows

- First regress Z on X, vis., $\hat{\Gamma} = \left(X'X\right)^{-1}\left(X'Z\right)$ and $\hat{Z} = X\hat{\Gamma} = P_X Z$.
- Second, regress Y on $\hat{Z}$, vis., $\hat{\beta} = \left(\hat{Z}'\hat{Z}\right)^{-1}\hat{Z}'Y$.

It is useful to scrutinize the projection $\hat{Z}$. Recall, $Z = [Z_1, Z_2]$ and $X = [Z_1, X_2]$. Then

$$\begin{aligned}\hat{Z} &= \left[\hat{Z}_1, \hat{Z}_2\right] \\ &= [P_X Z_1, P_X Z_2] \\ &= [Z_1, P_X Z_2] \\ &= \left[Z_1, \hat{Z}_2\right],\end{aligned}$$

since Z_1 lies in the span of X. Thus in the second stage, we regress Y on Z_1 and $\hat{Z}_2$. So only the endogenous variables Z_2 are replaced by their fitted values:

$$\hat{Z}_2 = X_1\hat{\Gamma}_{12} + X_2\hat{\Gamma}_{22}.$$

11.6 Bekker Asymptotics

Bekker (1994) used an alternative asymptotic framework to analyze the finite-sample bias in the 2SLS estimator. Here we present a simplified version of one of his results. In our notation, the model is

$$\begin{aligned}Y &= Z\beta + e \qquad (11.11)\\ Z &= X\Gamma + u \qquad (11.12)\\ \xi &= (e, u) \\ E(\xi \mid X) &= 0 \\ E\left(\xi\xi' \mid X\right) &= S\end{aligned}$$

As before, X is $n \times l$ so there are l instruments.

First, let's analyze the approximate bias of OLS applied to (11.11). Using (11.12),

$$E\left(\frac{1}{n}Z'e\right) = E(z_i e_i) = \Gamma' E(x_i e_i) + E(u_i e_i) = S_{21}$$

and

$$\begin{aligned}E\left(\frac{1}{n}Z'Z\right) &= E\left(z_i z_i'\right) \\ &= \Gamma' E\left(x_i x_i'\right)\Gamma + E\left(u_i x_i'\right)\Gamma + \Gamma' E\left(x_i u_i'\right) + E\left(u_i u_i'\right) \\ &= \Gamma' Q \Gamma + S_{22}\end{aligned}$$

where $Q = E\left(x_i x_i'\right)$. Hence by a first-order approximation

$$\begin{aligned}E\left(\hat{\beta}_{OLS} - \beta\right) &\approx \left(E\left(\frac{1}{n}Z'Z\right)\right)^{-1} E\left(\frac{1}{n}Z'e\right) \\ &= \left(\Gamma' Q \Gamma + S_{22}\right)^{-1} S_{21} \qquad (11.13)\end{aligned}$$

which is zero only when $S_{21} = 0$ (when Z is exogenous).

We now derive a similar result for the 2SLS estimator.

$$\hat{\beta}_{2SLS} = \left(Z' P_X Z\right)^{-1}\left(Z' P_X Y\right).$$

Let $P_X = X\left(X'X\right)^{-1}X'$. By the spectral decomposition of an idempotent matrix, $P = H\Lambda H'$ where $\Lambda = diag(I_l, 0)$. Let $q = H'\xi S^{-1/2}$ which satisfies $Eqq' = I_n$ and partition $q = (q_1'\ q_2')$ where q_1 is $l \times 1$. Hence

$$\begin{aligned} E\left(\frac{1}{n}\xi' P_X \xi\right) &= \frac{1}{n}S^{1/2\prime}E\left(q'\Lambda q\right)S^{1/2} \\ &= \frac{1}{n}S^{1/2\prime}E\left(\frac{1}{n}q_1'q_1\right)S^{1/2} \\ &= \frac{l}{n}S^{1/2\prime}S^{1/2} \\ &= \alpha S \end{aligned}$$

where

$$\alpha = \frac{l}{n}.$$

Using (11.12) and this result,

$$\frac{1}{n}E\left(Z'P_X e\right) = \frac{1}{n}E\left(\Gamma'X'e\right) + \frac{1}{n}E\left(u'P_X e\right) = \alpha S_{21},$$

and

$$\begin{aligned} \frac{1}{n}E\left(Z'P_X Z\right) &= \Gamma'E\left(x_i x_i'\right)\Gamma + \Gamma'E\left(x_i u_i\right) + E\left(u_i x_i'\right)\Gamma + \frac{1}{n}E\left(u'P_X u\right) \\ &= \Gamma'Q\Gamma + \alpha S_{22}. \end{aligned}$$

Together

$$\begin{aligned} E\left(\hat{\beta}_{2SLS} - \beta\right) &\approx \left(E\left(\frac{1}{n}Z'P_X Z\right)\right)^{-1} E\left(\frac{1}{n}Z'P_X e\right) \\ &= \alpha\left(\Gamma'Q\Gamma + \alpha S_{22}\right)^{-1} S_{21}. \end{aligned} \tag{11.14}$$

In general this is non-zero, except when $S_{21} = 0$ (when Z is exogenous). It is also close to zero when $\alpha = 0$. Bekker (1994) pointed out that it also has the reverse implication – that when $\alpha = l/n$ is large, the bias in the 2SLS estimator will be large. Indeed as $\alpha \to 1$, the expression in (11.14) approaches that in (11.13), indicating that the bias in 2SLS approaches that of OLS as the number of instruments increases.

Bekker (1994) showed further that under the alternative asymptotic approximation that α is fixed as $n \to \infty$ (so that the number of instruments goes to infinity proportionately with sample size) then the expression in (11.14) is the probability limit of $\hat{\beta}_{2SLS} - \beta$

11.7 Identification Failure

Recall the reduced form equation

$$Z_2 = X_1\Gamma_{12} + X_2\Gamma_{22} + u_2.$$

The parameter β fails to be identified if Γ_{22} has deficient rank. The consequences of identification failure for inference are quite severe.

Take the simplest case where $k = l = 1$ (so there is no X_1). Then the model may be written as

$$\begin{aligned} y_i &= z_i\beta + e_i \\ z_i &= x_i\gamma + u_i \end{aligned}$$

and $\Gamma_{22} = \gamma = E(x_i z_i)/Ex_i^2$. We see that β is identified if and only if $\Gamma_{22} = \gamma \neq 0$, which occurs when $E(z_i x_i) \neq 0$. Thus identification hinges on the existence of correlation between the excluded exogenous variable and the included endogenous variable.

Suppose this condition fails, so $E(z_i x_i) = 0$. Then by the CLT

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i e_i \to_d N_1 \sim N\left(0, E\left(x_i^2 e_i^2\right)\right) \tag{11.15}$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i z_i = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i u_i \to_d N_2 \sim N\left(0, E\left(x_i^2 u_i^2\right)\right) \tag{11.16}$$

therefore

$$\hat{\beta} - \beta = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i e_i}{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i z_i} \to_d \frac{N_1}{N_2} \sim Cauchy,$$

since the ratio of two normals is Cauchy. This is particularly nasty, as the Cauchy distribution does not have a finite mean. This result carries over to more general settings, and was examined by Phillips (1989) and Choi and Phillips (1992).

Suppose that identification does not complete fail, but is *weak*. This occurs when Γ_{22} is full rank, but *small*. This can be handled in an asymptotic analysis by modeling it as local-to-zero, viz

$$\Gamma_{22} = n^{-1/2}C,$$

where C is a full rank matrix. The $n^{-1/2}$ is picked because it provides just the right balancing to allow a rich distribution theory.

To see the consequences, once again take the simple case $k = l = 1$. Here, the instrument x_i is weak for z_i if

$$\gamma = n^{-1/2}c.$$

Then (11.15) is unaffected, but (11.16) instead takes the form

$$\begin{aligned} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i z_i &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i^2\gamma + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i u_i \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 c + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i u_i \\ &\to_d Qc + N_2 \end{aligned}$$

therefore

$$\hat{\beta} - \beta \to_d \frac{N_1}{Qc + N_2}.$$

As in the case of complete identification failure, we find that $\hat{\beta}$ is inconsistent for β and the asymptotic distribution of $\hat{\beta}$ is non-normal. In addition, standard test statistics have non-standard distributions, meaning that inferences about parameters of interest can be misleading.

The distribution theory for this model was developed by Staiger and Stock (1997) and extended to nonlinear GMM estimation by Stock and Wright (2000). Further results on testing were obtained by Wang and Zivot (1998).

The bottom line is that it is highly desirable to avoid identification failure. Once again, the equation to focus on is the reduced form

$$Z_2 = X_1\Gamma_{12} + X_2\Gamma_{22} + u_2$$

and identification requires $rank(\Gamma_{22}) = k_2$. If $k_2 = 1$, this requires $\Gamma_{22} \neq 0$, which is straightforward to assess using a hypothesis test on the reduced form. Therefore in the case of $k_2 = 1$ (one RHS endogenous variable), one constructive recommendation is to explicitly estimate the reduced form equation for Z_2, construct the test of $\Gamma_{22} = 0$, and at a minimum check that the test rejects $H_0 : \Gamma_{22} = 0$.

When $k_2 > 1$, $\Gamma_{22} \neq 0$ is not sufficient for identification. It is not even sufficient that each column of Γ_{22} is non-zero (each column corresponds to a distinct endogenous variable in X_2). So while a minimal check is to test that each columns of Γ_{22} is non-zero, this cannot be interpreted as definitive proof that Γ_{22} has full rank. Unfortunately, tests of deficient rank are difficult to implement. In any event, it appears reasonable to explicitly estimate and report the reduced form equations for X_2, and attempt to assess the likelihood that Γ_{22} has deficient rank.

11.8 Exercises

1. Consider the single equation model

$$y_i = z_i\beta + e_i,$$

where y_i and z_i are both real-valued (1×1). Let $\hat{\beta}$ denote the IV estimator of β using as an instrument a dummy variable d_i (takes only the values 0 and 1). Find a simple expression for the IV estimator in this context.

2. In the linear model

$$\begin{aligned} y_i &= z_i'\beta + e_i \\ E(e_i \mid z_i) &= 0 \end{aligned}$$

suppose $\sigma_i^2 = E\left(e_i^2 \mid z_i\right)$ is known. Show that the GLS estimator of β can be written as an IV estimator using some instrument x_i. (Find an expression for x_i.)

3. Take the linear model

$$Y = Z\beta + e.$$

Let the OLS estimator for β be $\hat{\beta}$ and the OLS residual be $\hat{e} = Y - Z\hat{\beta}$.

Let the IV estimator for β using some instrument X be $\tilde{\beta}$ and the IV residual be $\tilde{e} = Y - Z\tilde{\beta}$. If X is indeed endogeneous, will IV "fit" better than OLS, in the sense that $\tilde{e}'\tilde{e} < \hat{e}'\hat{e}$, at least in large samples?

4. The reduced form between the regressors z_i and instruments x_i takes the form

$$z_i = x_i'\Gamma + u_i$$

or

$$Z = X\Gamma + u$$

where z_i is $k \times 1$, x_i is $l \times 1$, Z is $n \times k$, X is $n \times l$, u is $n \times k$, and Γ is $l \times k$. The parameter Γ is defined by the population moment condition

$$E\left(x_i u_i'\right) = 0$$

Show that the method of moments estimator for Γ is $\hat{\Gamma} = \left(X'X\right)^{-1}\left(X'Z\right)$.

5. In the structural model

$$\begin{aligned} Y &= Z\beta + e \\ Z &= X\Gamma + u \end{aligned}$$

with Γ $l \times k$, $l \geq k$, we claim that β is identified (can be recovered from the reduced form) if $rank(\Gamma) = k$. Explain why this is true. That is, show that if $rank(\Gamma) < k$ then β cannot be identified.

6. Take the linear model

$$\begin{aligned} y_i &= x_i\beta + e_i \\ E(e_i \mid x_i) &= 0. \end{aligned}$$

where x_i and β are 1×1.

(a) Show that $E(x_i e_i) = 0$ and $E\left(x_i^2 e_i\right) = 0$. Is $z_i = (x_i \quad x_i^2)$ a valid instrumental variable for estimation of β?

(b) Define the 2SLS estimator of β, using z_i as an instrument for x_i. How does this differ from OLS?

(c) Find the efficient GMM estimator of β based on the moment condition

$$E(z_i(y_i - x_i\beta)) = 0.$$

Does this differ from 2SLS and/or OLS?

7. Suppose that price and quantity are determined by the intersection of the linear demand and supply curves

$$\begin{aligned} \text{Demand} \quad &: \quad Q = a_0 + a_1 P + a_2 Y + e_1 \\ \text{Supply} \quad &: \quad Q = b_0 + b_1 P + b_2 W + e_2 \end{aligned}$$

where income (Y) and wage (W) are determined outside the market. In this model, are the parameters identified?

8. The data file *card.dat* is taken from David Card "Using Geographic Variation in College Proximity to Estimate the Return to Schooling" in *Aspects of Labour Market Behavior* (1995). There are 2215 observations with 19 variables, listed in card.pdf. We want to estimate a wage equation

$$\log(Wage) = \beta_0 + \beta_1 Educ + \beta_2 Exper + \beta_3 Exper^2 + \beta_4 South + \beta_5 Black + e$$

where $Educ = Eduation$ (Years) $Exper = Experience$ (Years), and $South$ and $Black$ are regional and racial dummy variables.

(a) Estimate the model by OLS. Report estimates and standard errors.

(b) Now treat *Education* as endogenous, and the remaining variables as exogenous. Estimate the model by 2SLS, using the instrument *near*4, a dummy indicating that the observation lives near a 4-year college. Report estimates and standard errors.

(c) Re-estimate by 2SLS (report estimates and standard errors) adding three additional instruments: *near2* (a dummy indicating that the observation lives near a 2-year college), $fatheduc$ (the education, in years, of the father) and *motheduc* (the education, in years, of the mother).

(d) Re-estimate the model by efficient GMM. I suggest that you use the 2SLS estimates as the first-step to get the weight matrix, and then calculate the GMM estimator from this weight matrix without further iteration. Report the estimates and standard errors.

(e) Calculate and report the J statistic for overidentification.

(f) Discuss your findings..

Chapter 12

Univariate Time Series

A time series y_t is a process observed in sequence over time, $t = 1, ..., T$. To indicate the dependence on time, we adopt new notation, and use the subscript t to denote the individual observation, and T to denote the number of observations.

Because of the sequential nature of time series, we expect that Y_t and Y_{t-1} are *not* independent, so classical assumptions are not valid.

We can separate time series into two categories: univariate ($y_t \in R$ is scalar); and multivariate ($y_t \in R^m$ is vector-valued). The primary model for univariate time series is autoregressions (ARs). The primary model for multivariate time series is vector autoregressions (VARs).

12.1 Stationarity and Ergodicity

Definition 12.1.1 *$\{Y_t\}$ is covariance (weakly) stationary if*

$$E(Y_t) = \mu$$

is independent of t, and

$$Cov\,(Y_t, Y_{t-k}) = \gamma(k)$$

is independent of t for all k.

$\gamma(k)$ is called the autocovariance function.

Definition 12.1.2 *$\{Y_t\}$ is strictly stationary if the joint distribution of $(Y_t, ..., Y_{t-k})$ is independent of t for all k.*

Definition 12.1.3 *$\rho(k) = \gamma(k)/\gamma(0) = Corr(Y_t, Y_{t-k})$ is the autocorrelation function.*

Definition 12.1.4 *(loose). A stationary time series is ergodic if $\gamma(k) \to 0$ as $k \to \infty$.*

The following two theorems are essential to the analysis of stationary time series. There proofs are rather difficult, however.

Theorem 12.1.1 *If Y_t is strictly stationary and ergodic and $X_t = f(Y_t, Y_{t-1}, ...)$ is a random variable, then X_t is strictly stationary and ergodic.*

Theorem 12.1.2 *(Ergodic Theorem). If X_t is strictly stationary and ergodic and $E|X_t| < \infty$, then as $T \to \infty$,*

$$\frac{1}{T}\sum_{t=1}^{T} X_t \to_p E(X_t).$$

This allows us to consistently estimate parameters using time-series moments:
The sample mean:

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} Y_t$$

The sample autocovariance

$$\hat{\gamma}(k) = \frac{1}{T}\sum_{t=1}^{T} (Y_t - \hat{\mu})(Y_{t-k} - \hat{\mu}).$$

The sample autocorrelation

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}.$$

Theorem 12.1.3 *If Y_t is strictly stationary and ergodic and $EY_t^2 < \infty$, then as $T \to \infty$,*

1. $\hat{\mu} \to_p E(Y_t)$;
2. $\hat{\gamma}(k) \to_p \gamma(k)$;
3. $\hat{\rho}(k) \to_p \rho(k)$.

Proof. Part (1) is a direct consequence of the Ergodic theorem. For Part (2), note that

$$\begin{aligned}\hat{\gamma}(k) &= \frac{1}{T}\sum_{t=1}^{T} (Y_t - \hat{\mu})(Y_{t-k} - \hat{\mu}) \\ &= \frac{1}{T}\sum_{t=1}^{T} Y_t Y_{t-k} - \frac{1}{T}\sum_{t=1}^{T} Y_t\hat{\mu} - \frac{1}{T}\sum_{t=1}^{T} Y_{t-k}\hat{\mu} + \hat{\mu}^2.\end{aligned}$$

By Theorem 12.1.1 above, the sequence Y_tY_{t-k} is strictly stationary and ergodic, and it has a finite mean by the assumption that $EY_t^2 < \infty$. Thus an application of the Ergodic Theorem yields

$$\frac{1}{T}\sum_{t=1}^{T} Y_t Y_{t-k} \to_p E(Y_t Y_{t-k}).$$

Thus

$$\hat{\gamma}(k) \to_p E(Y_tY_{t-k}) - \mu^2 - \mu^2 + \mu^2 = E(Y_tY_{t-k}) - \mu^2 = \gamma(k).$$

Part (3) follows by the continuous mapping theorem: $\hat{\rho}(k) = \hat{\gamma}(k)/\hat{\gamma}(0) \to_p \gamma(k)/\gamma(0) = \rho(k)$. ■

12.2 Autoregressions

In time-series, the series $\{..., Y_1, Y_2, ..., Y_T, ...\}$ are jointly random. We consider the conditional expectation

$$E(Y_t \mid I_{t-1})$$

where $I_{t-1} = \{Y_{t-1}, Y_{t-2}, ...\}$ is the past history of the series.

An autoregressive (AR) model specifies that only a finite number of past lags matter:

$$E(Y_t \mid I_{t-1}) = E(Y_t \mid Y_{t-1}, ..., Y_{t-k}).$$

A linear AR model (the most common type used in practice) specifies linearity:

$$E(Y_t \mid I_{t-1}) = \alpha + \rho_1 Y_{t-1} + \rho_2 Y_{t-1} + \cdots + \rho_k Y_{t-k}.$$

Letting

$$e_t = Y_t - E(Y_t \mid I_{t-1}),$$

then we have the autoregressive model

$$\begin{aligned} Y_t &= \alpha + \rho_1 Y_{t-1} + \rho_2 Y_{t-1} + \cdots + \rho_k Y_{t-k} + e_t \\ E(e_t \mid I_{t-1}) &= 0. \end{aligned}$$

The last property defines a special time-series process.

Definition 12.2.1 *e_t is a martingale difference sequence (MDS) if $E(e_t \mid I_{t-1}) = 0$.*

Regression errors are naturally a MDS. Some time-series processes may be a MDS as a consequence of optimizing behavior. For example, some versions of the life-cycle hypothesis imply that either changes in consumption, or consumption growth rates, should be a MDS. Most asset pricing models imply that asset returns should be the sum of a constant plus a MDS.

The MDS property for the regression error plays the same role in a time-series regression as does the conditional mean-zero property for the regression error in a cross-section regression. In fact, it is even more important in the time-series context, as it is difficult to derive distribution theories without this property.

A useful property of a MDS is that e_t is uncorrelated with any function of the lagged information I_{t-1}. Thus for $k > 0$, $E(Y_{t-k} e_t) = 0$.

12.3 Stationarity of AR(1) Process

A mean-zero AR(1) is

$$Y_t = \rho Y_{t-1} + e_t.$$

Assume that e_t is iid, $E(e_t) = 0$ and $Ee_t^2 = \sigma^2 < \infty$.

By back-substitution, we find

$$\begin{aligned} Y_t &= e_t + \rho e_{t-1} + \rho^2 e_{t-2} + ... \\ &= \sum_{k=0}^{\infty} \rho^k e_{t-k}. \end{aligned}$$

Loosely speaking, this series converges if the sequence $\rho^k e_{t-k}$ gets small as $k \to \infty$. This occurs when $|\rho| < 1$.

Theorem 12.3.1 *If $|\rho| < 1$ then Y_t is strictly stationary and ergodic.*

We can compute the moments of Y_t using the infinite sum:

$$EY_t = \sum_{k=0}^{\infty} \rho^k E\left(e_{t-k}\right) = 0$$

$$Var(Y_t) = \sum_{k=0}^{\infty} \rho^{2k} Var\left(e_{t-k}\right) = \frac{\sigma^2}{1-\rho^2}.$$

If the equation for Y_t has an intercept, the above results are unchanged, except that the mean of Y_t can be computed from the relationship

$$EY_t = \alpha + \rho EY_{t-1},$$

and solving for $EY_t = EY_{t-1}$ we find $EY_t = \alpha/(1-\rho)$.

12.4 Lag Operator

An algebraic construct which is useful for the analysis of autoregressive models is the lag operator.

Definition 12.4.1 *The lag operator L satisfies $LY_t = Y_{t-1}$.*

Defining $L^2 = LL$, we see that $L^2 Y_t = LY_{t-1} = Y_{t-2}$. In general, $L^k Y_t = Y_{t-k}$.

The AR(1) model can be written in the format

$$Y_t - \rho Y_{t-1} + e_t$$

or

$$\left(1 - \rho L\right) Y_{t-1} = e_t.$$

The operator $\rho(L) = (1 - \rho L)$ is a polynomial in the operator L. We say that the *root* of the polynomial is $1/\rho$, since $\rho(z) = 0$ when $z = 1/\rho$. We call $\rho(L)$ the autoregressive polynomial of Y_t.

From Theorem 12.3.1, an AR(1) is stationary iff $|\rho| < 1$. Note that an equivalent way to say this is that an AR(1) is stationary iff the root of the autoregressive polynomial is larger than one (in absolute value).

12.5 Stationarity of AR(k)

The AR(k) model is

$$Y_t = \rho_1 Y_{t-1} + \rho_2 Y_{t-1} + \cdots + \rho_k Y_{t-k} + e_t.$$

Using the lag operator,

$$Y_t - \rho_1 LY_t - \rho_2 L^2 Y_t - \cdots - \rho_k L^k Y_t = e_t,$$

or

$$\rho(L) Y_t = e_t$$

where

$$\rho(L) = 1 - \rho_1 L - \rho_2 L^2 - \cdots - \rho_k L^k.$$

We call $\rho(L)$ the autoregressive polynomial of Y_t.

The *Fundamental Theorem of Algebra* says that any polynomial can be factored as

$$\rho(z) = \left(1 - \lambda_1^{-1} z\right)\left(1 - \lambda_2^{-1} z\right) \cdots \left(1 - \lambda_k^{-1} z\right)$$

where the $\lambda_1, ..., \lambda_k$ are the complex *roots* of $\rho(z)$, which satisfy $\rho(\lambda_j) = 0$.

We know that an AR(1) is stationary iff the absolute value of the root of its autoregressive polynomial is larger than one. For an AR(k), the requirement is that all roots are larger than one. Let $|\lambda|$ denote the modulus of a complex number λ.

Theorem 12.5.1 *The AR(k) is strictly stationary and ergodic if and only if $|\lambda_j| > 1$ for all j.*

One way of stating this is that "All roots lie outside the unit circle."

If one of the roots equals 1, we say that $\rho(L)$, and hence Y_t, "has a unit root". This is a special case of non-stationarity, and is of great interest in applied time series.

12.6 Estimation

Let

$$\begin{aligned} x_t &= \begin{pmatrix} 1 & Y_{t-1} & Y_{t-2} & \cdots & Y_{t-k} \end{pmatrix}' \\ \beta &= \begin{pmatrix} \alpha & \rho_1 & \rho_2 & \cdots & \rho_k \end{pmatrix}'. \end{aligned}$$

Then the model can be written as

$$y_t = x_t'\beta + e_t.$$

The OLS estimator is

$$\hat{\beta} = \left(X'X\right)^{-1} X'Y.$$

To study $\hat{\beta}$, it is helpful to define the process $u_t = x_t e_t$. Note that u_t is a MDS, since

$$E\left(u_t \mid I_{t-1}\right) = E\left(x_t e_t \mid I_{t-1}\right) = x_t E\left(e_t \mid I_{t-1}\right) = 0.$$

By Theorem 12.1.1, it is also strictly stationary and ergodic. Thus

$$\frac{1}{T}\sum_{t=1}^{T} x_t e_t = \frac{1}{T}\sum_{t=1}^{T} u_t \to_p E\left(u_t\right) = 0. \tag{12.1}$$

Theorem 12.6.1 *If the AR(k) process Y_t is strictly stationary and ergodic and $EY_t^2 < \infty$, then $\hat{\beta} \to_p \beta$ as $T \to \infty$.*

Proof. The vector x_t is strictly stationary and ergodic, and by Theorem 12.1.1, so is $x_t x_t'$. Thus by the Ergodic Theorem,

$$\frac{1}{T}\sum_{t=1}^{T} x_t x_t' \to_p E\left(x_t x_t'\right) = Q.$$

Combined with (12.1) and the continuous mapping theorem, we see that

$$\hat{\beta} = \beta + \left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} x_t e_t\right) \to_p Q^{-1} 0 = 0.$$

■

12.7 Asymptotic Distribution

Theorem 12.7.1 *MDS CLT. If u_t is a strictly stationary and ergodic MDS and $E(u_t u_t') = \Omega < \infty$, then as $T \to \infty$,*

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} u_t \to_d N(0, \Omega).$$

Since $x_t e_t$ is a MDS, we can apply Theorem 12.7.1 to see that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t e_t \to_d N(0, \Omega),$$

where

$$\Omega = E(x_t x_t' e_t^2).$$

Theorem 12.7.2 *If the AR(k) process Y_t is strictly stationary and ergodic and $EY_t^4 < \infty$, then as $T \to \infty$,*

$$\sqrt{T}\left(\hat{\beta} - \beta\right) \to_d N\left(0, Q^{-1}\Omega Q^{-1}\right).$$

This is identical in form to the asymptotic distribution of OLS in cross-section regression. The implication is that asymptotic inference is the same. In particular, the asymptotic covariance matrix is estimated just as in the cross-section case.

12.8 Bootstrap for Autoregressions

In the non-parametric bootstrap, we constructed the bootstrap sample by randomly resampling from the data values $\{y_t, x_t\}$. This creates an iid bootstrap sample. Clearly, this cannot work in a time-series application, as this imposes inappropriate independence.

Briefly, there are two popular methods to implement bootstrap resampling for time-series data.

Method 1: Model-Based (Parametric) Bootstrap.

1. Estimate $\hat{\beta}$ and residuals $\hat{e}_t$.
2. Fix an initial condition $(Y_{-k+1}, Y_{-k+2}, ..., Y_0)$.
3. Simulate iid draws e_i^* from the empirical distribution of the residuals $\{\hat{e}_1, ..., \hat{e}_T\}$.
4. Create the bootstrap series Y_t^* by the recursive formula

$$Y_t^* = \hat{\alpha} + \hat{\rho}_1 Y_{t-1}^* + \hat{\rho}_2 Y_{t-2}^* + \cdots + \hat{\rho}_k Y_{t-k}^* + e_t^*.$$

This construction imposes homoskedasticity on the errors e_i^*, which may be different than the properties of the actual e_i. It also presumes that the AR(k) structure is the truth.

Method 2: Block Resampling

1. Divide the sample into T/m blocks of length m.

2. Resample complete blocks. For each simulated sample, draw T/m blocks.

3. Paste the blocks together to create the bootstrap time-series Y_t^*.

4. This allows for arbitrary stationary serial correlation, heteroskedasticity, and for model-misspecification.

5. The results may be sensitive to the block length, and the way that the data are partitioned into blocks.

6. May not work well in small samples.

12.9 Trend Stationarity

$$\begin{aligned} Y_t &= \mu_0 + \mu_1 t + S_t \quad (12.2) \\ S_t &= \rho_1 S_{t-1} + \rho_2 S_{t-2} + \cdots + \rho_k S_{t-l} + e_t, \quad (12.3) \end{aligned}$$

or

$$Y_t = \alpha_0 + \alpha_1 t + \rho_1 Y_{t-1} + \rho_2 Y_{t-1} + \cdots + \rho_k Y_{t-k} + e_t. \quad (12.4)$$

There are two essentially equivalent ways to estimate the autoregressive parameters $(\rho_1, ..., \rho_k)$.

- You can estimate (12.4) by OLS.
- You can estimate (12.2)-(12.3) sequentially by OLS. That is, first estimate (12.2), get the residual $\hat{S}_t$, and then perform regression (12.3) replacing S_t with $\hat{S}_t$. This procedure is sometimes called *Detrending*.

The reason why these two procedures are (essentially) the same is the Frisch-Waugh-Lovell theorem.

Seasonal Effects

There are three popular methods to deal with seasonal data.

- Include dummy variables for each season. This presumes that "seasonality" does not change over the sample.
- Use "seasonally adjusted" data. The seasonal factor is typically estimated by a two-sided weighted average of the data for that season in neighboring years. Thus the seasonally adjusted data is a "filtered" series. This is a flexible approach which can extract a wide range of seasonal factors. The seasonal adjustment, however, also alters the time-series correlations of the data.
- First apply a seasonal differencing operator. If s is the number of seasons (typically $s = 4$ or $s = 12$),

  $$\Delta_s Y_t = Y_t - Y_{t-s},$$

  or the season-to-season change. The series $\Delta_s Y_t$ is clearly free of seasonality. But the long-run trend is also eliminated, and perhaps this was of relevance.

12.10 Testing for Omitted Serial Correlation

For simplicity, let the null hypothesis be an AR(1):

$$Y_t = \alpha + \rho Y_{t-1} + u_t. \tag{12.5}$$

We are interested in the question if the error u_t is serially correlated. We model this as an AR(1):

$$u_t = \theta u_{t-1} + e_t \tag{12.6}$$

with e_t a MDS. The hypothesis of no omitted serial correlation is

$$\begin{aligned} H_0 &: \theta = 0 \\ H_1 &: \theta \neq 0. \end{aligned}$$

We want to test H_0 against H_1.

To combine (12.5) and (12.6), we take (12.5) and lag the equation once:

$$Y_{t-1} = \alpha + \rho Y_{t-2} + u_{t-1}.$$

We then multiply this by θ and subtract from (12.5), to find

$$Y_t - \theta Y_{t-1} = \alpha - \theta\alpha + \rho Y_{t-1} - \theta\rho Y_{t-1} + u_t - \theta u_{t-1},$$

or

$$Y_t = \alpha(1-\theta) + (\rho + \theta) Y_{t-1} - \theta\rho Y_{t-2} + e_t = AR(2).$$

Thus under H_0, Y_t is an AR(1), and under H_1 it is an AR(2). H_0 may be expressed as the restriction that the coefficient on Y_{t-2} is zero.

An appropriate test of H_0 against H_1 is therefore a Wald test that the coefficient on Y_{t-2} is zero. (A simple exclusion test).

In general, if the null hypothesis is that Y_t is an AR(k), and the alternative is that the error is an AR(m), this is the same as saying that under the alternative Y_t is an AR(k+m), and this is equivalent to the restriction that the coefficients on $Y_{t-k-1}, ..., Y_{t-k-m}$ are jointly zero. An appropriate test is the Wald test of this restriction.

12.11 Model Selection

What is the appropriate choice of k in practice? This is a problem of model selection.

One approach to model selection is to choose k based on a Wald tests.

Another is to minimize the AIC or BIC information criterion, e.g.

$$AIC(k) = \log \hat{\sigma}^2(k) + \frac{2k}{T},$$

where $\hat{\sigma}^2(k)$ is the estimated residual variance from an AR(k)

One ambiguity in defining the AIC criterion is that the sample available for estimation changes as k changes. (If you increase k, you need more initial conditions.) This can induce strange behavior in the AIC. The best remedy is to fix a upper value $\overline{k}$, and then reserve the first $\overline{k}$ as initial conditions, and then estimate the models AR(1), AR(2), ..., AR($\overline{k}$) on this (unified) sample.

12.12 Autoregressive Unit Roots

The AR(k) model is

$$\begin{aligned} \rho(L)Y_t &= \mu + e_t \\ \rho(L) &= 1 - \rho_1 L - \cdots - \rho_k L^k. \end{aligned}$$

As we discussed before, Y_t has a unit root when $\rho(1) = 0$, or

$$\rho_1 + \rho_2 + \cdots + \rho_k = 1.$$

In this case, Y_t is non-stationary. The ergodic theorem and MDS CLT do not apply, and test statistics are asymptotically non-normal.

A helpful way to write the equation is the so-called Dickey-Fuller reparameterization:

$$\Delta Y_t = \mu + \alpha_0 Y_{t-1} + \alpha_1 \Delta Y_{t-1} + \cdots + \alpha_{k-1} \Delta Y_{t-(k-1)} + e_t. \tag{12.7}$$

These models are equivalent linear transformations of one another. The DF parameterization is convenient because the parameter α_0 summarizes the information about the unit root, since $\rho(1) = -\alpha_0$. To see this, observe that the lag polynomial for the Y_t computed from (12.7) is

$$(1 - L) - \alpha_0 L - \alpha_1 (L - L^2) - \cdots - \alpha_{k-1}(L^{k-1} - L^k)$$

But this must equal $\rho(L)$, as the models are equivalent. Thus

$$\rho(1) = (1 - 1) - \alpha_0 - (1 - 1) - \cdots - (1 - 1) = -\alpha_0.$$

Hence, the hypothesis of a unit root in Y_t can be stated as

$$H_0 : \alpha_0 = 0.$$

Note that the model is stationary if $\alpha_0 < 0$. So the natural alternative is

$$H_1 : \alpha_0 < 0.$$

Under H_0, the model for Y_t is

$$\Delta Y_t = \mu + \alpha_1 \Delta Y_{t-1} + \cdots + \alpha_{k-1} \Delta Y_{t-(k-1)} + e_t,$$

which is an AR(k-1) in the first-difference ΔY_t. Thus if Y_t has a (single) unit root, then ΔY_t is a stationary AR process. Because of this property, we say that if Y_t is non-stationary but $\Delta^d Y_t$ is stationary, then Y_t is "integrated of order d", or $I(d)$. Thus a time series with unit root is $I(1)$.

Since α_0 is the parameter of a linear regression, the natural test statistic is the t-statistic for H_0 from OLS estimation of (12.7). Indeed, this is the most popular unit root test, and is called the Augmented Dickey-Fuller (ADF) test for a unit root.

It would seem natural to assess the significance of the ADF statistic using the normal table. However, under H_0, Y_t is non-stationary, so conventional normal asymptotics are invalid. An alternative asymptotic framework has been developed to deal with non-stationary data. We do not have the time to develop this theory in detail, but simply assert the main results.

Theorem 12.12.1 *(Dickey-Fuller Theorem). Assume* $\alpha_0 = 0$. *As* $T \to \infty$,

$$T\hat{\alpha}_0 \to_d (1 - \alpha_1 - \alpha_2 - \cdots - \alpha_{k-1})\, DF_\alpha$$

$$ADF = \frac{\hat{\alpha}_0}{s(\hat{\alpha}_0)} \to DF_t.$$

The limit distributions DF_α and DF_t are non-normal. They are skewed to the left, and have negative means.

The first result states that $\hat{\alpha}_0$ converges to its true value (of zero) at rate T, rather than the conventional rate of $T^{1/2}$. This is called a "super-consistent" rate of convergence.

The second result states that the t-statistic for $\hat{\alpha}_0$ converges to a limit distribution which is non-normal, but does not depend on the parameters α. This distribution has been extensively tabulated, and may be used for testing the hypothesis H_0. Note: The standard error $s(\hat{\alpha}_0)$ is the conventional ("homoskedastic") standard error. But the theorem does not require an assumption of homoskedasticity. Thus the Dickey-Fuller test is robust to heteroskedasticity.

Since the alternative hypothesis is one-sided, the ADF test rejects H_0 in favor of H_1 when $ADF < c$, where c is the critical value from the ADF table. If the test rejects H_0, this means that the evidence points to Y_t being stationary. If the test does not reject H_0, a common conclusion is that the data suggests that Y_t is non-stationary. This is not really a correct conclusion, however. All we can say is that there is insufficient evidence to conclude whether the data are stationary or not.

We have described the test for the setting of with an intercept. Another popular setting includes as well a linear time trend. This model is

$$\Delta Y_t = \mu_1 + \mu_2 t + \alpha_0 Y_{t-1} + \alpha_1 \Delta Y_{t-1} + \cdots + \alpha_{k-1} \Delta Y_{t-(k-1)} + e_t. \tag{12.8}$$

This is natural when the alternative hypothesis is that the series is stationary about a linear time trend. If the series has a linear trend (e.g. GDP, Stock Prices), then the series itself is non-stationary, but it may be stationary around the linear time trend. In this context, it is a silly waste of time to fit an AR model to the level of the series without a time trend, as the AR model cannot conceivably describe this data. The natural solution is to include a time trend in the fitted OLS equation. When conducting the ADF test, this means that it is computed as the t-ratio for α_0 from OLS estimation of (12.8).

If a time trend is included, the test procedure is the same, but different critical values are required. The ADF test has a different distribution when the time trend has been included, and a different table should be consulted.

Most texts include as well the critical values for the extreme polar case where the intercept has been omitted from the model. These are included for completeness (from a pedagogical perspective) but have no relevance for empirical practice where intercepts are always included.

Chapter 13

Multivariate Time Series

A multivariate time series Y_t is a vector process $m \times 1$. Let $I_{t-1} = (Y_{t-1}, Y_{t-2}, ...)$ be all lagged information at time t. The typical goal is to find the conditional expectation $E(Y_t \mid I_{t-1})$. Note that since Y_t is a vector, this conditional expectation is also a vector.

13.1 Vector Autoregressions (VARs)

A VAR model specifies that the conditional mean is a function of only a finite number of lags:

$$E(Y_t \mid I_{t-1}) = E(Y_t \mid Y_{t-1}, ..., Y_{t-k}).$$

A linear VAR specifies that this conditional mean is linear in the arguments:

$$E(Y_t \mid Y_{t-1}, ..., Y_{t-k}) = A_0 + A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots A_k Y_{t-k}.$$

Observe that A_0 is $m \times 1$,and each of A_1 through A_k are $m \times m$ matrices.

Defining the $m \times 1$ regression error

$$e_t = Y_t - E(Y_t \mid I_{t-1}),$$

we have the VAR model

$$\begin{aligned} Y_t &= A_0 + A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots A_k Y_{t-k} + e_t \\ E(e_t \mid I_{t-1}) &= 0. \end{aligned}$$

Alternatively, defining the $mk + 1$ vector

$$x_t = \begin{pmatrix} 1 \\ Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-k} \end{pmatrix}$$

and the $m \times (mk + 1)$ matrix

$$A = \begin{pmatrix} A_0 & A_1 & A_2 & \cdots & A_k \end{pmatrix},$$

then

$$Y_t = Ax_t + e_t.$$

The VAR model is a system of m equations. One way to write this is to let a_j' be the jth row of A. Then the VAR system can be written as the equations

$$Y_{jt} = a_j' x_t + e_{jt}.$$

Unrestricted VARs were introduced to econometrics by Sims (1980).

13.2 Estimation

Consider the moment conditions

$$E\left(x_t e_{jt}\right) = 0,$$

$j = 1, ..., m$. These are implied by the VAR model, either as a regression, or as a linear projection.

The GMM estimator corresponding to these moment conditions is equation-by-equation OLS

$$\hat{a}_j = (X'X)^{-1} X'Y_j.$$

An alternative way to compute this is as follows. Note that

$$\hat{a}_j' = Y_j' X (X'X)^{-1}.$$

And if we stack these to create the estimate $\hat{A}$, we find

$$\begin{aligned} \hat{A} &= \begin{pmatrix} Y_1' \\ Y_2' \\ \vdots \\ Y_{m+1}' \end{pmatrix} X (X'X)^{-1} \\ &= Y'X(X'X)^{-1}, \end{aligned}$$

where

$$Y = \begin{pmatrix} Y_1 & Y_2 & \cdots & Y_m \end{pmatrix}$$

the $T \times m$ matrix of the stacked y_t'.

This (system) estimator is known as the SUR (Seemingly Unrelated Regressions) estimator, and was originally derived by Zellner (1962)

13.3 Restricted VARs

The unrestricted VAR is a system of m equations, each with the same set of regressors. A restricted VAR imposes restrictions on the system. For example, some regressors may be excluded from some of the equations. Restrictions may be imposed on individual equations, or across equations. The GMM framework gives a convenient method to impose such restrictions on estimation.

13.4 Single Equation from a VAR

Often, we are only interested in a single equation out of a VAR system. This takes the form

$$Y_{jt} = a_j' x_t + e_t,$$

and x_t consists of lagged values of Y_{jt} and the other $Y_{lt}'s$. In this case, it is convenient to re-define the variables. Let $y_t = Y_{jt}$, and Z_t be the other variables. Let $e_t = e_{jt}$ and $\beta = a_j$. Then the single equation takes the form

$$y_t = x_t'\beta + e_t, \tag{13.1}$$

and

$$x_t = \left[\begin{pmatrix} 1 & Y_{t-1} & \cdots & Y_{t-k} & Z_{t-1}' & \cdots & Z_{t-k}' \end{pmatrix}'\right].$$

This is just a conventional regression, with time series data.

13.5 Testing for Omitted Serial Correlation

Consider the problem of testing for omitted serial correlation in equation (13.1). Suppose that e_t is an AR(1). Then

$$\begin{aligned} y_t &= x_t'\beta + e_t \\ e_t &= \theta e_{t-1} + u_t \\ E\left(u_t \mid I_{t-1}\right) &= 0. \end{aligned} \tag{13.2}$$

Then the null and alternative are

$$H_0 : \theta = 0 \qquad H_1 : \theta \neq 0.$$

Take the equation $y_t = x_t'\beta + e_t$, and subtract off the equation once lagged multiplied by θ, to get

$$\begin{aligned} y_t - \theta y_{t-1} &= \left(x_t'\beta + e_t\right) - \theta\left(x_{t-1}'\beta + e_{t-1}\right) \\ &= x_t'\beta - \theta x_{t-1}\beta + e_t - \theta e_{t-1}, \end{aligned}$$

or

$$y_t = \theta y_{t-1} + x_t'\beta + x_{t-1}'\gamma + u_t, \tag{13.3}$$

which is a valid regression model.

So testing H_0 versus H_1 is equivalent to testing for the significance of adding (y_{t-1}, x_{t-1}) to the regression. This can be done by a Wald test. We see that an appropriate, general, and simple way to test for omitted serial correlation is to test the significance of extra lagged values of the dependent variable and regressors.

You may have heard of the Durbin-Watson test for omitted serial correlation, which once was very popular, and is still routinely reported by conventional regression packages. The DW test is appropriate only when regression $y_t = x_t'\beta + e_t$ is not dynamic (has no lagged values on the RHS), and e_t is iid $N(0,1)$. Otherwise it is invalid.

Another interesting fact is that (13.2) is a special case of (13.3), under the restriction $\gamma = -\beta\theta$. This restriction, which is called a common factor restriction, may be tested if desired. If valid, the model (13.2) may be estimated by iterated GLS. (A simple version of this estimator is called Cochrane-Orcutt.) Since the common factor restriction appears arbitrary, and is typically rejected empirically, direct estimation of (13.2) is uncommon in recent applications.

13.6 Selection of Lag Length in an VAR

If you want a data-dependent rule to pick the lag length k in a VAR, you may either use a testing-based approach (using, for example, the Wald statistic), or an information criterion approach. The formula for the AIC and BIC are

$$\begin{aligned} AIC(k) &= \log\det\left(\hat{\Omega}(k)\right) + 2\frac{p}{T} \\ BIC(k) &= \log\det\left(\hat{\Omega}(k)\right) + \frac{p\log(T)}{T} \\ \hat{\Omega}(k) &= \frac{1}{T}\sum_{t=1}^{T}\hat{e}_t(k)\hat{e}_t(k)' \\ p &= m(km+1) \end{aligned}$$

where p is the number of parameters in the model, and $\hat{e}_t(k)$ is the OLS residual vector from the model with k lags. The log determinant is the criterion from the multivariate normal likelihood.

13.7 Granger Causality

Partition the data vector into (Y_t, Z_t). Define the two information sets

$$\begin{aligned} I_{1t} &= (Y_t, Y_{t-1}, Y_{t-2}, ...) \\ I_{2t} &= (Y_t, Z_t, Y_{t-1}, Z_{t-1}, Y_{t-2}, Z_{t-2}, , ...) \end{aligned}$$

The information set I_{1t} is generated only by the history of Y_t, and the information set I_{2t} is generated by both Y_t and Z_t. The latter has more information.

We say that Z_t does not *Granger-cause* Y_t if

$$E\left(Y_t \mid I_{1,t-1}\right) = E\left(Y_t \mid I_{2,t-1}\right).$$

That is, conditional on information in lagged Y_t, lagged Z_t does not help to forecast Y_t. If this condition does not hold, then we say that Z_t Granger-causes Y_t.

The reason why we call this "Granger Causality" rather than "causality" is because this is not a physical or structure definition of causality. If Z_t is some sort of forecast of the future, such as a futures price, then Z_t may help to forecast Y_t even though it does not "cause" Y_t. This definition of causality was developed by Granger (1969) and Sims (1972).

In a linear VAR, the equation for Y_t is

$$Y_t = \alpha + \rho_1 Y_{t-1} + \cdots + \rho_k Y_{t-k} + Z_{t-1}'\gamma_1 + \cdots + Z_{t-k}'\gamma_k + e_t.$$

In this equation, Z_t does not Granger-cause Y_t if and only if

$$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_k = 0.$$

This may be tested using an exclusion (Wald) test.

This idea can be applied to blocks of variables. That is, Y_t and/or Z_t can be vectors. The hypothesis can be tested by using the appropriate multivariate Wald test.

If it is found that Z_t does not Granger-cause Y_t, then we deduce that our time-series model of $E\left(Y_t \mid I_{t-1}\right)$ does not require the use of Z_t. Note, however, that Z_t may still be useful to explain other features of Y_t, such as the conditional variance.

13.8 Cointegration

The idea of cointegration is due to Granger (1981), and was articulated in detail by Engle and Granger (1987).

Definition 13.8.1 *The $m \times 1$ series Y_t is cointegrated if Y_t is $I(1)$ yet there exists β, $m \times r$, of rank r, such that $z_t = \beta' Y_t$ is $I(0)$. The r vectors in β are called the cointegrating vectors.*

If the series Y_t is not cointegrated, then $r = 0$. If $r = m$, then Y_t is $I(0)$. For $0 < r < m$, Y_t is $I(1)$ and cointegrated.

In some cases, it may be believed that β is known a priori. Often, $\beta = (1 \quad -1)'$. For example, if Y_t is a pair of interest rates, then $\beta = (1 \quad -1)'$ specifies that the spread (the difference in returns) is stationary. If $Y = (\log(Consumption) \quad \log(Income))'$, then $\beta = (1 \quad -1)'$ specifies that $\log(Consumption/Income)$ is stationary.

In other cases, β may not be known.

If Y_t is cointegrated with a single cointegrating vector ($r = 1$), then it turns out that β can be consistently estimated by an OLS regression of one component of Y_t on the others. Thus $Y_t = (Y_{1t}, Y_{2t})$ and $\beta = (\beta_1 \ \beta_2)$ and normalize $\beta_1 = 1$. Then $\hat{\beta}_2 = (Y_2'Y_2)^{-1} Y_2 Y_1 \to_p \beta_2$. Furthermore this estimation is super-consistent: $T(\hat{\beta}_2 - \beta_2) \to_d Limit$, as first shown by Stock (1987). This is not, in general, a good method to estimate β, but it is useful in the construction of alternative estimators and tests.

We are often interested in testing the hypothesis of no cointegration:

$$\begin{aligned} H_0 &: r = 0 \\ H_1 &: r > 0. \end{aligned}$$

Suppose that β is known, so $z_t = \beta' Y_t$ is known. Then under H_0 z_t is $I(1)$, yet under H_1 z_t is $I(0)$. Thus H_0 can be tested using a univariate ADF test on z_t.

When β is unknown, Engle and Granger (1987) suggested using an ADF test on the estimated residual $\hat{z}_t = \hat{\beta}' Y_t$, from OLS of Y_{1t} on Y_{2t}. Their justification was Stock's result that $\hat{\beta}$ is super-consistent under H_1. Under H_0, however, $\hat{\beta}$ is not consistent, so the ADF critical values are not appropriate. The asymptotic distribution was worked out by Phillips and Ouliaris (1990).

When the data have time trends, it may be necessary to include a time trend in the estimated cointegrating regression. Whether or not the time trend is included, the asymptotic distribution of the test is affected by the presence of the time trend. The asymptotic distribution was worked out in B. Hansen (1992).

13.9 Cointegrated VARs

We can write a VAR as

$$\begin{aligned} A(L) Y_t &= e_t \\ A(L) &= I - A_1 L - A_2 L^2 - \cdots - A_k L^k \end{aligned}$$

or alternatively as

$$\Delta Y_t = \Pi Y_{t-1} + D(L) \Delta Y_{t-1} + e_t$$

where

$$\begin{aligned} \Pi &= -A(1) \\ &= -I + A_1 + A_2 + \cdots + A_k. \end{aligned}$$

Theorem 13.9.1 *(Granger Representation Theorem). Y_t is cointegrated with $m \times r$ β if and only if $rank(\Pi) = r$ and $\Pi = \alpha\beta'$ where α is $m \times r$, $rank(\alpha) = r$.*

Thus cointegration imposes a restriction upon the parameters of a VAR. The restricted model can be written as

$$\begin{aligned} \Delta Y_t &= \alpha\beta' Y_{t-1} + D(L)\Delta Y_{t-1} + e_t \\ \Delta Y_t &= \alpha z_{t-1} + D(L)\Delta Y_{t-1} + e_t. \end{aligned}$$

If β is known, this can be estimated by OLS of ΔY_t on z_{t-1} and the lags of ΔY_t.

If β is unknown, then estimation is done by "reduced rank regression", which is least-squares subject to the stated restriction. Equivalently, this is the MLE of the restricted parameters under the assumption that e_t is iid $N(0, \Omega)$.

One difficulty is that β is not identified without normalization. When $r = 1$, we typically just normalize one element to equal unity. When $r > 1$, this does not work, and different authors have adopted different identification schemes.

In the context of a cointegrated VAR estimated by reduced rank regression, it is simple to test for cointegration by testing the rank of Π. These tests are constructed as likelihood ratio (LR) tests. As they were discovered by Johansen (1988, 1991, 1995), they are typically called the "Johansen Max and Trace" tests. Their asymptotic distributions are non-standard, and are similar to the Dickey-Fuller distributions.

Chapter 14

Limited Dependent Variables

A "limited dependent variable" Y is one which takes a "limited" set of values. The most common cases are

- Binary: $Y = \{0, 1\}$
- Multinomial: $Y = \{0, 1, 2, ..., k\}$
- Integer: $Y = \{0, 1, 2, ...\}$
- Censored: $Y = \{x : x \geq 0\}$

The traditional approach to the estimation of limited dependent variable (LDV) models is parametric maximum likelihood. A parametric model is constructed, allowing the construction of the likelihood function. A more modern approach is semi-parametric, eliminating the dependence on a parametric distributional assumption. We will discuss only the first (parametric) approach, due to time constraints. They still constitute the majority of LDV applications. If, however, you were to write a thesis involving LDV estimation, you would be advised to consider employing a semi-parametric estimation approach.

For the parametric approach, estimation is by MLE. A major practical issue is construction of the likelihood function.

14.1 Binary Choice

The dependent variable $Y_i = \{0, 1\}$. This represents a Yes/No outcome. Given some regressors x_i, the goal is to describe $P(Y_i = 1 \mid x_i)$, as this is the full conditional distribution.

The linear probability model specifies that

$$P(Y_i = 1 \mid x_i) = x_i'\beta.$$

As $P(Y_i = 1 \mid x_i) = E(Y_i \mid x_i)$, this yields the regression: $Y_i = x_i'\beta + e_i$ which can be estimated by OLS. However, the linear probability model does not impose the restriction that $0 \leq P(Y_i \mid x_i) \leq 1$. Even so estimation of a linear probability model is a useful starting point for subsequent analysis.

The standard alternative is to use a function of the form

$$P(Y_i = 1 \mid x_i) = F\left(x_i'\beta\right)$$

where $F(\cdot)$ is a known CDF, typically assumed to be symmetric about zero, so that $F(z) = 1 - F(-z)$. The two standard choices for F are

- Logistic: $F(u) = (1 + e^{-u})^{-1}$.
- Normal: $F(u) = \Phi(u)$.

If F is logistic, we call this the *logit* model, and if F is normal, we call this the *probit* model. This model is identical to the latent variable model

$$\begin{aligned} Y_i^* &= x_i'\beta + e_i \\ e_i &\sim F(\cdot) \\ Y_i &= \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{otherwise} \end{cases}. \end{aligned}$$

For then

$$\begin{aligned} P(Y_i = 1 \mid x_i) &= P(Y_i^* > 0 \mid x_i) \\ &= P\left(x_i'\beta + e_i > 0 \mid x_i\right) \\ &= P\left(e_i > -x_i'\beta \mid x_i\right) \\ &= 1 - F\left(-x_i'\beta\right) \\ &= F\left(x_i'\beta\right). \end{aligned}$$

Estimation is by maximum likelihood. To construct the likelihood, we need the conditional distribution of an individual observation. Recall that if Y is Bernoulli, such that $P(Y = 1) = p$ and $P(Y = 0) = 1 - p$, then we can write the density of Y as

$$f(y) = p^y(1-p)^{1-y}, \qquad y = 0, 1.$$

In the Binary choice model, Y_i is conditionally Bernoulli with $P(Y_i = 1 \mid x_i) = p_i = F(x_i'\beta)$. Thus the conditional density is

$$\begin{aligned} f(y_i \mid x_i) &= p_i^{y_i}(1-p_i)^{1-y_i} \\ &= F\left(x_i'\beta\right)^{y_i}\left(1 - F\left(x_i'\beta\right)\right)^{1-y_i}. \end{aligned}$$

Hence the log-likelihood function is

$$\begin{aligned} l_n(\beta) &= \sum_{i=1}^n \log f(y_i \mid x_i) \\ &= \sum_{i=1}^n \log\left(F\left(x_i'\beta\right)^{y_i}\left(1 - F\left(x_i'\beta\right)\right)^{1-y_i}\right) \\ &= \sum_{i=1}^n \left[y_i \log F\left(x_i'\beta\right) + (1 - y_i)\log(1 - F\left(x_i'\beta\right))\right] \\ &= \sum_{y_i=1} \log F\left(x_i'\beta\right) + \sum_{y_i=0} \log(1 - F\left(x_i'\beta\right)). \end{aligned}$$

The MLE $\hat{\beta}$ is the value of β which maximizes $l_n(\beta)$. Standard errors and test statistics are computed by asymptotic approximations. Details of such calculations are left to more advanced courses.

14.2 Count Data

If $Y = \{0, 1, 2, ...\}$, a typical approach is to employ *Poisson regression.* This model specifies that

$$\begin{aligned} P(Y_i = k \mid x_i) &= \frac{\exp(-\lambda_i)\lambda_i^k}{k!}, \qquad k = 0, 1, 2, ... \\ \lambda_i &= \exp(x_i'\beta). \end{aligned}$$

The conditional density is the Poisson with parameter λ_i. The functional form for λ_i has been picked to ensure that $\lambda_i > 0$.

The log-likelihood function is

$$l_n(\beta) = \sum_{i=1}^{n} \log f(y_i \mid x_i) = \sum_{i=1}^{n} \left(-\exp(x_i'\beta) + y_i x_i'\beta - \log(y_i!)\right).$$

The MLE is the value $\hat{\beta}$ which maximizes $l_n(\beta)$.

Since

$$E(Y_i \mid x_i) = \lambda_i = \exp(x_i'\beta)$$

is the conditional mean, this motivates the label Poisson "regression."

Also observe that the model implies that

$$Var(Y_i \mid x_i) = \lambda_i = \exp(x_i'\beta),$$

so the model imposes the restriction that the conditional mean and variance of Y_i are the same. This may be considered restrictive. A generalization is the negative binomial.

14.3 Censored Data

The idea of "censoring" is that some data above or below a threshold are mis-reported at the threshold. Thus the model is that there is some latent process y_i^* with unbounded support, but we observe only

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* \geq 0 \\ 0 & \text{if } y_i^* < 0 \end{cases}. \tag{14.1}$$

(This is written for the case of the threshold being zero, any known value can substitute.) The observed data y_i therefore come from a mixed continuous/discrete distribution.

Censored models are typically applied when the data set has a meaningful proportion (say 5% or higher) of data at the boundary of the sample support. The censoring process may be explicit in data collection, or it may be a by-product of economic constraints.

An example of a data collection censoring is top-coding of income. In surveys, incomes above a threshold are typically reported at the threshold.

The first censored regression model was developed by Tobin (1958) to explain consumption of durable goods. Tobin observed that for many households, the consumption level (purchases) in a particular period was zero. He proposed the latent variable model

$$\begin{aligned} y_i^* &= x_i'\beta + e_i \\ e_i &\sim iid\ N(0, \sigma^2) \end{aligned}$$

with the observed variable y_i generated by the censoring equation (14.1). This model (now called the Tobit) specifies that the latent (or ideal) value of consumption may be negative (the household

would prefer to sell than buy). All that is reported is that the household purchased zero units of the good.

The naive approach to estimate β is to regress y_i on x_i. This does not work because regression estimates $E(Y_i \mid x_i)$, not $E(Y_i^* \mid x_i) = x_i'\beta$, and the latter is of interest. Thus OLS will be biased for the parameter of interest β.

[Note: it is still possible to estimate $E(Y_i \mid x_i)$ by LS techniques. The Tobit framework postulates that this is not inherently interesting, that the parameter of β is defined by an alternative statistical structure.]

Consistent estimation will be achieved by the MLE. To construct the likelihood, observe that the probability of being censored is

$$\begin{aligned} P(y_i = 0 \mid x_i) &= P(y_i^* < 0 \mid x_i) \\ &= P\left(x_i'\beta + e_i < 0 \mid x_i\right) \\ &= P\left(\frac{e_i}{\sigma} < -\frac{x_i'\beta}{\sigma} \mid x_i\right) \\ &= \Phi\left(-\frac{x_i'\beta}{\sigma}\right). \end{aligned}$$

The conditional distribution function above zero is Gaussian:

$$P(y_i = y \mid x_i) = \int_0^y \sigma^{-1}\phi\left(\frac{z - x_i'\beta}{\sigma}\right) dz, \qquad y > 0.$$

Therefore, the density function can be written as

$$f(y \mid x_i) = \Phi\left(-\frac{x_i'\beta}{\sigma}\right)^{1(y=0)} \left[\sigma^{-1}\phi\left(\frac{z - x_i'\beta}{\sigma}\right)\right]^{1(y>0)},$$

where $1(\cdot)$ is the indicator function.

Hence the log-likelihood is a mixture of the probit and the normal:

$$\begin{aligned} l_n(\beta) &= \sum_{i=1}^n \log f(y_i \mid x_i) \\ &= \sum_{y_i=0} \log \Phi\left(-\frac{x_i'\beta}{\sigma}\right) + \sum_{y_i>0} \log\left[\sigma^{-1}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right]. \end{aligned}$$

The MLE is the value $\hat{\beta}$ which maximizes $l_n(\beta)$.

14.4 Sample Selection

The problem of sample selection arises when the sample is a non-random selection of potential observations. This occurs when the observed data is systematically different from the population of interest. For example, if you ask for volunteers for an experiment, and they wish to extrapolate the effects of the experiment on a general population, you should worry that the people who volunteer may be systematically different from the general population. This has great relevance for the evaluation of anti-poverty and job-training programs, where the goal is to assess the effect of "training" on the general population, not just on the volunteers.

A simple sample selection model can be written as the latent model

$$\begin{aligned} y_i &= x_i'\beta + e_{1i} \\ T_i &= 1\left(z_i'\gamma + e_{0i} > 0\right) \end{aligned}$$

where $1(\cdot)$ is the indicator function. The dependent variable y_i is observed if (and only if) $T_i = 1$. Else it is unobserved.

For example, y_i could be a wage, which can be observed only if a person is employed. The equation for T_i is an equation specifying the probability that the person is employed.

The model is often completed by specifying that the errors are jointly normal

$$\begin{pmatrix} e_{0i} \\ e_{1i} \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & \sigma^2 \end{pmatrix}\right).$$

It is presumed that we observe $\{x_i, z_i, T_i\}$ for all observations.

Under the normality assumption,

$$e_{1i} = \rho e_{0i} + v_i,$$

where v_i is independent of $e_{0i} \sim N(0,1)$. A useful fact about the standard normal distribution is that

$$E\left(e_{0i} \mid e_{0i} > -x\right) = \lambda(x) = \frac{\phi(x)}{\Phi(x)},$$

and the function $\lambda(x)$ is called the inverse Mills ratio.

The naive estimator of β is OLS regression of y_i on x_i for those observations for which y_i is available. The problem is that this is equivalent to conditioning on the event $\{T_i = 1\}$. However,

$$\begin{aligned} E\left(e_{1i} \mid T_i = 1, Z_i\right) &= E\left(e_{1i} \mid \{e_{0i} > -z_i'\gamma\}, Z_i\right) \\ &= \rho E\left(e_{0i} \mid \{e_{0i} > -z_i'\gamma\}, Z_i\right) + E\left(v_i \mid \{e_{0i} > -z_i'\gamma\}, Z_i\right) \\ &= \rho\lambda\left(z_i'\gamma\right), \end{aligned}$$

which is non-zero. Thus

$$e_{1i} = \rho\lambda\left(z_i'\gamma\right) + u_i,$$

where

$$E\left(u_i \mid T_i = 1, Z_i\right) = 0.$$

Hence

$$y_i = x_i'\beta + \rho\lambda\left(z_i'\gamma\right) + u_i \tag{14.2}$$

is a valid regression equation for the observations for which $T_i = 1$.

Heckman (1979) observed that we could consistently estimate β and ρ from this equation, if γ were known. It is unknown, but also can be consistently estimated by a Probit model for selection. The "Heckit" estimator is thus calculated as follows

- Estimate $\hat{\gamma}$ from a Probit, using regressors z_i. The binary dependent variable is T_i.
- Estimate $\left(\hat{\beta}, \hat{\rho}\right)$ from OLS of y_i on x_i and $\lambda(z_i'\hat{\gamma})$.
- The OLS standard errors will be incorrect, as this is a two-step estimator. They can be corrected using a more complicated formula. Or, alternatively, by viewing the Probit/OLS estimation equations as a large joint GMM problem.

The Heckit estimator is frequently used to deal with problems of sample selection. However, the estimator is built on the assumption of normality, and the estimator can be quite sensitive to this assumption. Some modern econometric research is exploring how to relax the normality assumption.

The estimator can also work quite poorly if $\lambda\left(z_i'\hat{\gamma}\right)$ does not have much in-sample variation. This can happen if the Probit equation does not "explain" much about the selection choice. Another potential problem is that if $z_i = x_i$, then $\lambda\left(z_i'\hat{\gamma}\right)$ can be highly collinear with x_i, so the second step OLS estimator will not be able to precisely estimate β. Based this observation, it is typically recommended to find a valid exclusion restriction: a variable should be in z_i which is not in x_i. If this is valid, it will ensure that $\lambda\left(z_i'\hat{\gamma}\right)$ is not collinear with x_i, and hence improve the second stage estimator's precision.

Chapter 15

Panel Data

A panel is a set of observations on individuals, collected over time. An observation is the pair $\{y_{it}, x_{it}\}$, where the i subscript denotes the individual, and the t subscript denotes time. A panel may be *balanced*:

$$\{y_{it}, x_{it}\} : t = 1, ..., T; \quad i = 1, ..., n,$$

or *unbalanced*:

$$\{y_{it}, x_{it}\} : \text{For } i = 1, ..., n, \quad t = \underline{t}_i, ..., \overline{t}_i.$$

15.1 Individual-Effects Model

The standard panel data specification is that there is an individual-specific effect which enters linearly in the regression

$$y_{it} = x_{it}'\beta + u_i + e_{it}.$$

The typical maintained assumptions are that the individuals i are mutually independent, that u_i and e_{it} are independent, that e_{it} is iid across individuals and time, and that e_{it} is uncorrelated with x_{it}.

OLS of y_{it} on x_{it} is called pooled estimation. It is consistent if

$$E(x_{it}u_i) = 0 \tag{15.1}$$

If this condition fails, then OLS is inconsistent. (15.1) fails if the individual-specific unobserved effect u_i is correlated with the observed explanatory variables x_{it}. This is often believed to be plausible if u_i is an omitted variable.

If (15.1) is true, however, OLS can be improved upon via a GLS technique. In either event, OLS appears a poor estimation choice.

Condition (15.1) is called the *random effects hypothesis.* It is a strong assumption, and most applied researchers try to avoid its use.

15.2 Fixed Effects

This is the most common technique for estimation of non-dynamic linear panel regressions.

The motivation is to allow u_i to be arbitrary, and have arbitrary correlated with x_i. The goal is to eliminate u_i from the estimator, and thus achieve invariance.

There are several derivations of the estimator.

First, let

$$d_{ij} = \begin{cases} 1 & \text{if } i = j \\ \\ 0 & \text{else} \end{cases},$$

and

$$d_i = \begin{pmatrix} d_{i1} \\ \vdots \\ d_{in} \end{pmatrix},$$

an $n \times 1$ dummy vector with a "1" in the $i'th$ place. Let

$$u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}.$$

Then note that

$$u_i = d_i' u,$$

and

$$y_{it} = x_{it}'\beta + d_i' u + e_{it}. \tag{15.2}$$

Observe that

$$E\left(e_{it} \mid x_{it}, d_i\right) = 0,$$

so (15.2) is a valid regression, with d_i as a regressor along with x_i.

OLS on (15.2) yields estimator $(\hat{\beta}, \hat{u})$. Conventional inference applies.

Observe that

- This is generally consistent.
- If x_{it} contains an intercept, it will be collinear with d_i, so the intercept is typically omitted from x_{it}.
- Any regressor in x_{it} which is constant over time for all individuals (e.g., their gender) will be collinear with d_i, so will have to be omitted.
- There are $n + k$ regression parameters, which is quite large as typically n is very large.

Computationally, you do not want to actually implement conventional OLS estimation, as the parameter space is too large. OLS estimation of β proceeds by the FWL theorem. Stacking the observations together:

$$Y = X\beta + Du + e,$$

then by the FWL theorem,

$$\begin{aligned} \hat{\beta} &= \left(X'\left(1 - P_D\right)X\right)^{-1}\left(X'\left(1 - P_D\right)Y\right) \\ &= \left(X^{*\prime}X^*\right)^{-1}\left(X^{*\prime}Y^*\right), \end{aligned}$$

where

$$\begin{aligned} Y^* &= Y - D(D'D)^{-1}D'Y \\ X^* &= X - D(D'D)^{-1}D'X. \end{aligned}$$

Since the regression of y_{it} on d_i is a regression onto individual-specific dummies, the predicted value from these regressions is the individual specific mean $\overline{y}_i$, and the residual is the demean value

$$y_{it}^* = y_{it} - \overline{y}_i.$$

The fixed effects estimator $\hat{\beta}$ is OLS of y_{it}^* on x_{it}^*, the dependent variable and regressors in deviation-from-mean form.

Another derivation of the estimator is to take the equation

$$y_{it} = x_{it}'\beta + u_i + e_{it},$$

and then take individual-specific means by taking the average for the $i'th$ individual:

$$\frac{1}{T_i}\sum_{t=\underline{t}_i}^{\overline{t}_i} y_{it} = \frac{1}{T_i}\sum_{t=\underline{t}_i}^{\overline{t}_i} x_{it}'\beta + u_i + \frac{1}{T_i}\sum_{t=\underline{t}_i}^{\overline{t}_i} e_{it}$$

or

$$\overline{y}_i = \overline{x}_i'\beta + u_i + \overline{e}_i.$$

Subtracting, we find

$$y_{it}^* = x_{it}^{*\prime}\beta + e_{it}^*,$$

which is free of the individual-effect u_i.

15.3 Dynamic Panel Regression

A dynamic panel regression has a lagged dependent variable

$$y_{it} = \alpha y_{it-1} + x_{it}'\beta + u_i + e_{it}. \tag{15.3}$$

This is a model suitable for studying dynamic behavior of individual agents.

Unfortunately, the fixed effects estimator is inconsistent, at least if T is held finite as $n \to \infty$. This is because the sample mean of y_{it-1} is correlated with that of e_{it}.

The standard approach to estimate a dynamic panel is to combine first-differencing with IV or GMM. Taking first-differences of (15.3) eliminates the individual-specific effect:

$$\Delta y_{it} = \alpha \Delta y_{it-1} + \Delta x_{it}'\beta + \Delta e_{it}. \tag{15.4}$$

However, if e_{it} is iid, then it will be correlated with Δy_{it-1} :

$$E\left(\Delta y_{it-1}\Delta e_{it}\right) = E\left(\left(y_{it-1} - y_{it-2}\right)\left(e_{it} - e_{it-1}\right)\right) = -E\left(y_{it-1}e_{it-1}\right) = -\sigma_e^2.$$

So OLS on (15.4) will be inconsistent.

But if there are valid instruments, then IV or GMM can be used to estimate the equation. Typically, we use lags of the dependent variable, two periods back, as y_{t-2} is uncorrelated with Δe_{it}. Thus values of y_{it-k}, $k \geq 2$, are valid instruments.

Hence a valid estimator of α and β is to estimate (15.4) by IV using y_{t-2} as an instrument for Δy_{t-1} (which is just identified). Alternatively, GMM using y_{t-2} and y_{t-3} as instruments (which is overidentified, but loses a time-series observation).

A more sophisticated GMM estimator recognizes that for time-periods later in the sample, there are more instruments available, so the instrument list should be different for each equation. This is conveniently organized by the GMM principle, as this enables the moments from the different time-periods to be stacked together to create a list of all the moment conditions. A simple application of GMM yields the parameter estimates and standard errors.

Chapter 16

Nonparametrics

16.1 Kernel Density Estimation

Let X be a random variable with continuous distribution $F(x)$ and density $f(x) = \frac{d}{dx}F(x)$. The goal is to estimate $f(x)$ from a random sample $(X_1, ..., X_n\}$ While $F(x)$ can be estimated by the EDF $\hat{F}(x) = n^{-1}\sum_{i=1}^{n} 1(X_i \leq x)$, we cannot define $\frac{d}{dx}\hat{F}(x)$ since $\hat{F}(x)$ is a step function. The standard **nonparametric** method to estimate $f(x)$ is based on **smoothing** using a kernel.

While we are typically interested in estimating the entire function $f(x)$, we can simply focus on the problem where x is a specific fixed number, and then see how the method generalizes to estimating the entire function.

Definition 1 *$K(u)$ is a **second-order kernel function** if it is a symmetric zero-mean density function.*

Three common choices for kernels include the **Gaussian**

$$K(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

the **Epanechnikov**

$$K(x) = \begin{array}{cc} \frac{3}{4}\left(1-x^2\right), & |x| \leq 1 \\ 0 & |x| > 1 \end{array}$$

and the **Biweight** or **Quartic**

$$K(x) = \begin{array}{cc} \frac{15}{16}\left(1-x^2\right)^2, & |x| \leq 1 \\ 0 & |x| > 1 \end{array}$$

In practice, the choice between these three rarely makes a meaningful difference in the estimates.

The kernel functions are used to smooth the data. The amount of smoothing is controlled by the **bandwidth** $h > 0$. Let

$$K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right).$$

be the kernel K rescaled by the bandwidth h. The kernel density estimator of $f(x)$ is

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x).$$

This estimator is the average of a set of weights. If a large number of the observations X_i are near x, then the weights are relatively large and $\hat{f}(x)$ is larger. Conversely, if only a few X_i are near x, then the weights are small and $\hat{f}(x)$ is small. The bandwidth h controls the meaning of "near".

Interestingly, $\hat{f}(x)$ is a valid density. That is, $\hat{f}(x) \geq 0$ for all x, and

$$\int_{-\infty}^{\infty} \hat{f}(x)dx = \int_{-\infty}^{\infty} \frac{1}{n}\sum_{i=1}^{n} K_h\left(X_i - x\right) dx = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} K_h\left(X_i - x\right) dx = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} K\left(u\right) du = 1$$

where the second-to-last equality makes the change-of-variables $u = (X_i - x)/h$.

We can also calculate the moments of the density $\hat{f}(x)$. The mean is

$$\begin{aligned}
\int_{-\infty}^{\infty} x\hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} xK_h\left(X_i - x\right) dx \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} \left(X_i + uh\right) K\left(u\right) du \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i \int_{-\infty}^{\infty} K\left(u\right) du + \frac{1}{n}\sum_{i=1}^{n} h \int_{-\infty}^{\infty} uK\left(u\right) du \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i
\end{aligned}$$

the sample mean of the X_i, where the second-to-last equality used the change-of-variables $u = (X_i - x)/h$ which has Jacobian h.

The second moment of the estimated density is

$$\begin{aligned}
\int_{-\infty}^{\infty} x^2\hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} x^2 K_h\left(X_i - x\right) dx \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} \left(X_i + uh\right)^2 K\left(u\right) du \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \frac{2}{n}\sum_{i=1}^{n} X_i h \int_{-\infty}^{\infty} K(u)du + \frac{1}{n}\sum_{i=1}^{n} h^2 \int_{-\infty}^{\infty} u^2 K\left(u\right) du \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + h^2\sigma_K^2
\end{aligned}$$

where

$$\sigma_K^2 = \int_{-\infty}^{\infty} x^2 K\left(x\right) dx$$

is the variance of the kernel. It follows that the variance of the density $\hat{f}(x)$ is

$$\begin{aligned}
\int_{-\infty}^{\infty} x^2\hat{f}(x)dx - \left(\int_{-\infty}^{\infty} x\hat{f}(x)dx\right)^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + h^2\sigma_K^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\
&= \hat{\sigma}^2 + h^2\sigma_K^2
\end{aligned}$$

Thus the variance of the estimated density is inflated by the factor $h^2\sigma_K^2$ relative to the sample moment.

16.2 Asymptotic MSE for Kernel Estimates

For fixed x and bandwidth h observe that

$$EK_h(X-x) = \int_{-\infty}^{\infty} K_h(z-x) f(z)dz = \int_{-\infty}^{\infty} K_h(uh) f(x+hu)hdu = \int_{-\infty}^{\infty} K(u) f(x+hu)du$$

The second equality uses the change-of variables $u = (z-x)/h$. The last expression shows that the expected value is an average of $f(z)$ locally about x.

This integral (typically) is not analytically solvable, so we approximate it using a second order Taylor expansion of $f(x+hu)$ in the argument hu about $hu = 0$, which is valid as $h \to 0$. Thus

$$f(x+hu) \simeq f(x) + f'(x)hu + \frac{1}{2}f''(x)h^2u^2$$

and therefore

$$\begin{aligned} EK_h(X-x) &\simeq \int_{-\infty}^{\infty} K(u)\left(f(x) + f'(x)hu + \frac{1}{2}f''(x)h^2u^2\right)du \\ &= f(x)\int_{-\infty}^{\infty} K(u)\,du + f'(x)h\int_{-\infty}^{\infty} K(u)\,udu + \frac{1}{2}f''(x)h^2\int_{-\infty}^{\infty} K(u)\,u^2du \\ &= f(x) + \frac{1}{2}f''(x)h^2\sigma_K^2. \end{aligned}$$

The bias of $\hat{f}(x)$ is then

$$Bias(x) = E\hat{f}(x) - f(x) = \frac{1}{n}\sum_{i=1}^{n} EK_h(X_i - x) - f(x) = \frac{1}{2}f''(x)h^2\sigma_K^2.$$

We see that the bias of $\hat{f}(x)$ at x depends on the second derivative $f''(x)$. The sharper the derivative, the greater the bias. Intuitively, the estimator $\hat{f}(x)$ smooths data local to $X_i = x$, so is estimating a smoothed version of $f(x)$. The bias results from this smoothing, and is larger the greater the curvature in $f(x)$.

We now examine the variance of $\hat{f}(x)$. Since it is an average of iid random variables, using first-order Taylor approximations and the fact that n^{-1} is of smaller order than $(nh)^{-1}$

$$\begin{aligned} Var(x) &= \frac{1}{n}Var(K_h(X_i - x)) \\ &= \frac{1}{n}EK_h(X_i - x)^2 - \frac{1}{n}(EK_h(X_i - x))^2 \\ &\simeq \frac{1}{nh^2}\int_{-\infty}^{\infty} K\left(\frac{z-x}{h}\right)^2 f(z)dz - \frac{1}{n}f(x)^2 \\ &= \frac{1}{nh}\int_{-\infty}^{\infty} K(u)^2 f(x+hu)\,du \\ &\simeq \frac{f(x)}{nh}\int_{-\infty}^{\infty} K(u)^2\,du \\ &= \frac{f(x)R(K)}{nh}. \end{aligned}$$

where $R(K) = \int_{-\infty}^{\infty} K(x)^2\,dx$ is called the **roughness** of K.

Together, the asymptotic mean-squared error (AMSE) for fixed x is the sum of the approximate squared bias and approximate variance

$$AMSE_h(x) = \frac{1}{4} f''(x)^2 h^4 \sigma_K^4 + \frac{f(x) R(K)}{nh}.$$

A global measure of precision is the asymptotic mean integrated squared error (AMISE)

$$AMISE_h = \int AMSE_h(x) dx = \frac{h^4 \sigma_K^4 R(f'')}{4} + \frac{R(K)}{nh}. \tag{16.1}$$

where $R(f'') = \int (f''(x))^2 dx$ is the roughness of f''. Notice that the first term (the squared bias) is increasing in h and the second term (the variance) is decreasing in nh. Thus for the AMISE to decline with n, we need $h \to 0$ but $nh \to \infty$. That is, h must tend to zero, but at a slower rate than n^{-1}.

Equation (16.1) is an asymptotic approximation to the MSE. We define the asymptotically optimal bandwidth h_0 as the value which minimizes this approximate MSE. That is,

$$h_0 = \underset{h}{\operatorname{argmin}} AMISE_h$$

It can be found by solving the first order condition

$$\frac{d}{dh} AMISE_h = h^3 \sigma_K^4 R(f'') - \frac{R(K)}{nh^2} = 0$$

yielding

$$h_0 = \left(\frac{R(K)}{\sigma_K^4 R(f'')} \right)^{1/5} n^{-1/2}. \tag{16.2}$$

This solution takes the form $h_0 = cn^{-1/5}$ where c is a function of K and f, but not of n. We thus say that the optimal bandwidth is of order $O(n^{-1/5})$. Note that this h declines to zero, but at a very slow rate.

In practice, how should the bandwidth be selected? This is a difficult problem, and there is a large and continuing literature on the subject. The asymptotically optimal choice given in (16.2) depends on $R(K)$, σ_K^2, and $R(f'')$. The first two are determined by the kernel function. Their values for the three functions introduced in the previous section are given here.

K	$\sigma_K^2 = \int_{-\infty}^{\infty} x^2 K(x) dx$	$R(K) = \int_{-\infty}^{\infty} K(x)^2 dx$
Gaussian	1	$1/(2\sqrt{\pi})$
Epanechnikov	1/5	1/5
Biweight	1/7	5/7

An obvious difficulty is that $R(f'')$ is unknown. A classic simple solution proposed by Silverman (1986)has come to be known as the **reference bandwidth** or **Silverman's Rule-of-Thumb**. It uses formula (16.2) but replaces $R(f'')$ with $\hat{\sigma}^{-5} R(\phi'')$, where ϕ is the $N(0,1)$ distribution and $\hat{\sigma}^2$ is an estimate of $\sigma^2 = Var(X)$. This choice for h gives an optimal rule when $f(x)$ is normal, and gives a nearly optimal rule when $f(x)$ is close to normal. The downside is that if the density is very far from normal, the rule-of-thumb h can be quite inefficient. We can calculate that $R(\phi'') = 3/(8\sqrt{\pi})$. Together with the above table, we find the reference rules for the three kernel functions introduced earlier.

Gaussian Kernel: $h_{rule} = 1.06 n^{-1/5}$

Epanechnikov Kernel: $h_{rule} = 2.34n^{-1/5}$
Biweight (Quartic) Kernel: $h_{rule} = 2.78n^{-1/5}$

Unless you delve more deeply into kernel estimation methods the rule-of-thumb bandwidth is a good practical bandwidth choice, perhaps adjusted by visual inspection of the resulting estimate $\hat{f}(x)$. There are other approaches, but implementation can be delicate. I now discuss some of these choices. The **plug-in** approach is to estimate $R(f'')$ in a first step, and then plug this estimate into the formula (16.2). This is more treacherous than may first appear, as the optimal h for estimation of the roughness $R(f'')$ is quite different than the optimal h for estimation of $f(x)$. However, there are modern versions of this estimator work well, in particular the iterative method of Sheather and Jones (1991). Another popular choice for selection of h is **cross-validation**. This works by constructing an estimate of the MISE using leave-one-out estimators. There are some desirable properties of cross-validation bandwidths, but they are also known to converge very slowly to the optimal values. They are also quite ill-behaved when the data has some discretization (as is common in economics), in which case the cross-validation rule can sometimes select very small bandwidths leading to dramatically undersmoothed estimates. Fortunately there are remedies, which are known as **smoothed cross-validation** which is a close cousin of the **bootstrap**.

Appendix A

Probability

A.1 Foundations

The set S of all possible outcomes of an experiment is called the **sample space** for the experiment. Take the simple example of tossing a coin. There are two outcomes, heads and tails, so we can write $S = \{H, T\}$. If two coins are tossed in sequence, we can write the four outcomes as $S = \{HH, HT, TH, TT\}$.

An **event** A is any collection of possible outcomes of an experiment. An event is a subset of S, including S itself and the null set $\emptyset$. Continuing the two coin example, one event is $A = \{HH, HT\}$, the event that the first coin is heads. We say that A and B are **disjoint** or **mutually exclusive** if $A \cap B = \emptyset$. For example, the sets $\{HH, HT\}$ and $\{TH\}$ are disjoint. Furthermore, if the sets $A_1, A_2, ...$ are pairwise disjoint and $\cup_{i=1}^{\infty} A_i = S$, then the collection $A_1, A_2, ...$ is called a **partition** of S.

The following are elementary set operations:

Union: $A \cup B = \{x : x \in A \text{ or } x \in B\}$.

Intersection: $A \cap B = \{x : x \in A \text{ and } x \in B\}$.

Complement: $A^c = \{x : x \notin A\}$.

The following are useful properties of set operations.

Communtatitivity: $A \cup B = B \cup A$; $\quad A \cap B = B \cap A$.

Associativity: $A \cup (B \cup C) = (A \cup B) \cup C$; $\quad A \cap (B \cap C) = (A \cap B) \cap C$.

Distributive Laws: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$; $\quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

DeMorgan's Laws: $(A \cup B)^c = A^c \cap B^c$; $\quad (A \cap B)^c = A^c \cup B^c$.

A **probability function** assigns probabilities (numbers between 0 and 1) to events A in S. This is straightforward when S is countable; when S is uncountable we must be somewhat more careful.A set $\mathcal{B}$ is called a **sigma algebra** (or Borel field) if $\emptyset \in \mathcal{B}$, $A \in \mathcal{B}$ implies $A^c \in \mathcal{B}$, and $A_1, A_2, ... \in \mathcal{B}$ implies $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$. A simple example is $\{\emptyset, S\}$ which is known as the trivial sigma algebra. For any sample space S, let $\mathcal{B}$ be the smallest sigma algebra which contains all of the open sets in S. When S is countable, $\mathcal{B}$ is simply the collection of all subsets of S, including $\emptyset$ and S. When S is the real line, then $\mathcal{B}$ is the collection of all open and closed intervals. We call $\mathcal{B}$ the sigma algebra associated with S. We only define probabilities for events contained in $\mathcal{B}$.

We now can give the axiomatic definition of probability. Given S and $\mathcal{B}$, a probability function P satisfies $P(S) = 1$, $P(A) \geq 0$ for all $A \in \mathcal{B}$, and if $A_1, A_2, ... \in \mathcal{B}$ are pairwise disjoint, then $P\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

Some important properties of the probability function include the following

- $P(\emptyset) = 0$

- $P(A) \leq 1$
- $P\left(A^c\right) = 1 - P(A)$
- $P\left(B \cap A^c\right) = P(B) - P(A \cap B)$
- $P\left(A \cup B\right) = P(A) + P(B) - P(A \cap B)$
- If $A \subset B$ then $P(A) \leq P(B)$
- Bonferroni's Inequality: $P(A \cap B) \geq P(A) + P(B) - 1$
- Boole's Inequality: $P\left(A \cup B\right) \leq P(A) + P(B)$

For some elementary probability models, it is useful to have simple rules to count the number of objects in a set. These counting rules are facilitated by using the binomial coefficients which are defined for nonnegative integers n and r, $n \geq r$, as

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!}.$$

When counting the number of objects in a set, there are two important distinctions. Counting may be **with replacement** or **without replacement**. Counting may be **ordered** or **unordered**. For example, consider a lottery where you pick six numbers from the set 1, 2, ..., 49. This selection is without replacement if you are not allowed to select the same number twice, and is with replacement if this is allowed. Counting is ordered or not depending on whether the sequential order of the numbers is relevant to winning the lottery. Depending on these two distinctions, we have four expressions for the number of objects (possible arrangements) of size r from n objects.

	Without Replacement	With Replacement
Ordered	$\frac{n!}{(n-r)!}$	n^r
Unordered	$\binom{n}{r}$	$\binom{n+r-1}{r}$

In the lottery example, if counting is unordered and without replacement, the number of potential combinations is $\binom{49}{6} = 13,983,816$.

If $P(B) > 0$ the **conditional probability** of the event A given the event B is

$$P\left(A \mid B\right) = \frac{P\left(A \cap B\right)}{P(B)}.$$

For any B, the conditional probability function is a valid probability function where S has been replaced by B. Rearranging the definition, we can write

$$P(A \cap B) = P\left(A \mid B\right) P(B)$$

which is often quite useful. We can say that the occurrence of B has no information about the likelihood of event A when $P\left(A \mid B\right) = P(A)$, in which case we find

$$P(A \cap B) = P\left(A\right) P(B) \tag{A.1}$$

We say that the events A and B are **statistically independent** when (A.1) holds. Furthermore, we say that the collection of events $A_1, ..., A_k$ are **mutually independent** when for any subset $\{A_i : i \in I\}$,

$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P\left(A_i\right).$$

Theorem 2 *(Bayes' Rule). For any set B and any partition $A_1, A_2, ...$ of the sample space, then for each $i = 1, 2, ...$*

$$P(A_i \mid B) = \frac{P(B \mid A_i) P(A_i)}{\sum_{j=1}^{\infty} P(B \mid A_j) P(A_j)}$$

A.2 Random Variables

A **random variable** X is a function from a sample space S into the real line. This induces a new sample space – the real line – and a new probability function on the real line. Typically, we denote random variables by uppercase letters such as X, and use lower case letters such as x for potential values and realized values. For a random variable X we define its **cumulative distribution function** (CDF) as

$$F(x) = P(X \leq x). \tag{A.2}$$

Sometimes we write this as $F_X(x)$ to denote that it is the CDF of X. A function $F(x)$ is a CDF if and only if the following three properties hold:

1. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$
2. $F(x)$ is nondecreasing in x
3. $F(x)$ is right-continuous

We say that the random variable X is **discrete** if $F(x)$ is a step function. In the latter case, the range of X consists of a countable set of real numbers $\tau_1, ..., \tau_r$. The probability function for X takes the form

$$P(X = \tau_j) = \pi_j, \qquad j = 1, ..., r \tag{A.3}$$

where $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{r} \pi_j = 1$.

We say that the random variable X is **continuous** if $F(x)$ is continuous in x. In this case $P(X = \tau) = 0$ for all $\tau \in R$ so the representation (A.3) is unavailable. Instead, we represent the relative probabilities by the **probability density function** (PDF)

$$f(x) = \frac{d}{dx} F(x)$$

so that

$$F(x) = \int_{-\infty}^{x} f(u) du$$

and

$$P(a \leq X \leq b) = \int_{a}^{b} f(x) dx.$$

These expressions only make sense if $F(x)$ is differentiable. While there are examples of continuous random variables which do not possess a PDF, these cases are unusual and are typically ignored.

A function $f(x)$ is a PDF if and only if $f(x) \geq 0$ for all $x \in R$ and $\int_{-\infty}^{\infty} f(x) dx$.

A.3 Expectation

For any measurable real function g, we define the **mean** or **expectation** $Eg(X)$ as follows. If X is discrete,

$$Eg(X) = \sum_{j=1}^{r} g(\tau_j)\pi_j,$$

and if X is continuous

$$Eg(X) = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

The latter is well defined and finite if

$$\int_{-\infty}^{\infty} |g(x)|\, f(x)dx < \infty. \tag{A.4}$$

If (A.4) does not hold, evaluate

$$\begin{aligned} I_1 &= \int_{g(x)>0} g(x)f(x)dx \\ I_2 &= -\int_{g(x)<0} g(x)f(x)dx \end{aligned}$$

If $I_1 = \infty$ and $I_2 < \infty$ then we define $Eg(X) = \infty$. If $I_1 < \infty$ and $I_2 = \infty$ then we define $Eg(X) = -\infty$. If both $I_1 = \infty$ and $I_2 = \infty$ then $Eg(X)$ is undefined.

Since $E(a + bX) = a + bEX$, we say that expectation is a linear operator.

For $m > 0$, we define the $m'th$ **moment** of X as EX^m and the $m'th$ **central moment** as $E(X - EX)^m$.

Two special moments are the **mean** $\mu = EX$ and **variance** $\sigma^2 = E(X - \mu)^2 = EX^2 - \mu^2$. We call $\sigma = \sqrt{\sigma^2}$ the **standard deviation** of X. We can also write $\sigma^2 = Var(X)$. For example, this allows the convenient expression $Var(a + bX) = b^2 Var(X)$.

The **moment generating function** (MGF) of X is

$$M(\lambda) = E\exp(\lambda X).$$

The MGF does not necessarily exist. However, when it does and $E|X|^m < \infty$ then

$$\frac{d^m}{d\lambda^m} M(\lambda)\bigg|_{\lambda=0} = E(X^m)$$

which is why it is called the moment generating function.

More generally, the **characteristic function** (CF) of X is

$$C(\lambda) = E\exp(i\lambda X).$$

where $i = \sqrt{-1}$ is the imaginary unit. The CF always exists, and when $E|X|^m < \infty$

$$\frac{d^m}{d\lambda^m} C(\lambda)\bigg|_{\lambda=0} = i^m E(X^m).$$

The L^p **norm**, $p \geq 1$, of the random variable X is

$$\|X\|_p = (E|X|^p)^{1/p}.$$

A.4 Common Distributions

For reference, we now list some important discrete distribution function.

Bernoulli

$$\begin{aligned} P(X=x) &= p^x(1-p)^{1-x}, \qquad x=0,1; \qquad 0\le p\le 1 \\ EX &= p \\ Var(X) &= p(1-p) \end{aligned}$$

Binomial

$$\begin{aligned} P(X=x) &= \binom{n}{x} p^x (1-p)^{n-x}, \qquad x=0,1,...,n; \qquad 0\le p\le 1 \\ EX &= np \\ Var(X) &= np(1-p) \end{aligned}$$

Geometric

$$\begin{aligned} P(X=x) &= p(1-p)^{x-1}, \qquad x=1,2,...; \qquad 0\le p\le 1 \\ EX &= \frac{1}{p} \\ Var(X) &= \frac{1-p}{p^2} \end{aligned}$$

Multinomial

$$\begin{aligned} P(X_1=x_1, X_2=x_2, ..., X_m=x_m) &= \frac{n!}{x_1!x_2!\cdots x_m!} p_1^{x_1} p_2^{x_2}\cdots p_m^{x_m}, \\ x_1+\cdots+x_m &= n; \\ p_1+\cdots+p_m &= 1 \\ EX &= \\ Var(X) &= \end{aligned}$$

Negative Binomial

$$\begin{aligned} P(X=x) &= \binom{r+x-1}{x} p(1-p)^{x-1}, \qquad x=1,2,...; \qquad 0\le p\le 1 \\ EX &= \\ Var(X) &= \end{aligned}$$

Poisson

$$\begin{aligned} P(X=x) &= \frac{\exp(-\lambda)\lambda^x}{x!}, \qquad x=0,1,2,..., \qquad \lambda>0 \\ EX &= \lambda \\ Var(X) &= \lambda \end{aligned}$$

We now list some important continuous distributions.

Beta

$$\begin{aligned} f(x) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 \leq x \leq 1; \qquad \alpha > 0,\ \beta > 0 \\ \mu &= \frac{\alpha}{\alpha+\beta} \\ Var(X) &= \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2} \end{aligned}$$

Cauchy

$$\begin{aligned} f(x) &= \frac{1}{\pi(1+x^2)}, \qquad -\infty < x < \infty \\ EX &= \infty \\ Var(X) &= \infty \end{aligned}$$

Exponential

$$\begin{aligned} f(x) &= \frac{1}{\theta}\exp\left(\frac{x}{\theta}\right), \qquad 0 \leq x < \infty; \qquad \theta > 0 \\ EX &= \theta \\ Var(X) &= \theta^2 \end{aligned}$$

Logistic

$$\begin{aligned} f(x) &= \frac{\exp(-x)}{(1+\exp(-x))^2}, \qquad -\infty < x < \infty; \\ EX &= 0 \\ Var(X) &= \end{aligned}$$

$\int_0^1 x^{-2}dx$

$$\int_0^\infty \frac{\exp(-x)}{(1+\exp(-x))^2}dx$$

Lognormal

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi}\sigma x}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right), \qquad 0 \leq x < \infty; \qquad \sigma > 0 \\ EX &= \exp\left(\mu+\sigma^2/2\right) \\ Var(X) &= \exp\left(2\mu+2\sigma^2\right) - \exp\left(2\mu+\sigma^2\right) \end{aligned}$$

Pareto

$$\begin{aligned} f(x) &= \frac{\beta\alpha^\beta}{x^{\beta+1}}, \qquad \alpha \leq x < \infty, \qquad \alpha > 0, \qquad \beta > 0 \\ EX &= \frac{\beta\alpha}{\beta-1}, \qquad \beta > 1 \\ Var(X) &= \frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)}, \qquad \beta > 2 \end{aligned}$$

Uniform

$$\begin{aligned} f(x) &= \frac{1}{b-a}, \qquad a \le x \le b \\ EX &= \frac{a+b}{2} \\ Var(X) &= \frac{(b-a)^2}{12} \end{aligned}$$

Weibull

$$\begin{aligned} f(x) &= \frac{\gamma}{\beta} x^{\gamma-1} \exp\left(-\frac{x^\gamma}{\beta}\right), \qquad 0 \le x < \infty; \qquad \gamma > 0,\ \beta > 0 \\ EX &= \beta^{1/\gamma} \Gamma\left(1 + \frac{1}{\gamma}\right) \\ Var(X) &= \beta^{2/\gamma} \left(\Gamma\left(1 + \frac{2}{\gamma}\right) - \Gamma^2\left(1 + \frac{1}{\gamma}\right)\right) \end{aligned}$$

A.5 Multivariate Random Variables

A pair of bivariate random variables (X, Y) is a function from the sample space into R^2. The joint CDF of (X, Y) is

$$F(x, y) = P\left(X \le x, Y \le y\right).$$

If F is continuous, the joint probability density function is

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y).$$

For a Borel measurable set $A \in R^2$,

$$P\left((X < Y) \in A\right) = \int \int_A f(x, y) dx dy$$

For any measurable function $g(x, y)$,

$$Eg(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy.$$

The **marginal distribution** of X is

$$\begin{aligned} F_X(x) &= P(X \le x) \\ &= \lim_{y \to \infty} F(x, y) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(x, y) dy dx \end{aligned}$$

so the **marginal density** of X is

$$f_X(x) = \frac{d}{dx} F_X(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$

Similarly, the marginal density of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx.$$

The random variables X and Y are defined to be **independent** if $f(x,y) = f_X(x)f_Y(y)$. Furthermore, X and Y are independent if and only if there exist functions $g(x)$ and $h(y)$ such that $f(x,y) = g(x)h(y)$.

If X and Y are independent, then

$$\begin{aligned} E\left(g(X)h(Y)\right) &= \int\int g(x)h(y)f(y,x)dydx \\ &= \int\int g(x)h(y)f_Y(y)f_X(x)dydx \\ &= \int g(x)f_X(x)dx \int h(y)f_Y(y)dy \\ &= Eg\left(X\right)Eh\left(Y\right). \end{aligned} \tag{A.5}$$

if the expectations exist. For example, if X and Y are independent then

$$E(XY) = EXEY.$$

Another implication of (A.5) is that if X and Y are independent and $Z = X + Y$, then

$$\begin{aligned} M_Z(\lambda) &= E\exp\left(\lambda\left(X+Y\right)\right) \\ &= E\left(\exp\left(\lambda X\right)\exp\left(\lambda Y\right)\right) \\ &= E\exp\left(\lambda' X\right)E\exp\left(\lambda' Y\right) \\ &= M_X(\lambda)M_Y(\lambda). \end{aligned} \tag{A.6}$$

The covariance between X and Y is

$$Cov(X,Y) = \sigma_{XY} = E\left(\left(X-EX\right)\left(Y-EY\right)\right) = EXY - EXEY.$$

The correlation between X and Y is

$$Corr(X,Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_x\sigma_Y}.$$

The Cauchy-Schwarz Inequality implies that $|\rho_{XY}| \leq 1$.The correlation is a measure of linear dependence, free of units of measurement.

If X and Y are independent, then $\sigma_{XY} = 0$ and $\rho_{XY} = 0$. The reverse, however, is not true. For example, if $EX = 0$ and $EX^3 = 0$, then $Cov(X, X^2) = 0$.

A useful fact is that

$$Var\left(X+Y\right) = Var(X) + Var(Y) + 2Cov(X,Y).$$

An implication is that if X and Y are independent, then

$$Var\left(X+Y\right) = Var(X) + Var(Y),$$

the variance of the sum is the sum of the variances.

A $k \times 1$ random vector $X = (X_1, ..., X_k)'$ is a function from S to R^k. Letting $x = (x_1, ..., x_k)'$, it has the distribution and density functions

$$\begin{aligned} F(x) &= P(X \leq x) \\ f(x) &= \frac{\partial^k}{\partial x_1 \cdots \partial x_k} F(x). \end{aligned}$$

For a measurable function $g : R^k \to R^s$, we define the expectation

$$Eg(X) = \int_{R^k} g(x) f(x) dx$$

where the symbol dx denotes $dx_1 \cdots dx_k$. In particular, we have the $k \times 1$ multivariate mean

$$\mu = EX$$

and $k \times k$ covariance matrix

$$\begin{aligned} \Sigma &= E\left((X - \mu)(X - \mu)'\right) \\ &= EXX' - \mu\mu' \end{aligned}$$

If the elements of X are mutually independent, then Σ is a diagonal matrix and

$$Var\left(\sum_{i=1}^{k} X_i\right) = \sum_{i=1}^{k} Var\left(X_i\right)$$

A.6 Conditional Distributions and Expectation

The **conditional density** of Y given $X = x$ is defined as

$$f_{Y|X}(y \mid x) = \frac{f(x, y)}{f_X(x)}$$

if $f_X(x) > 0$. One way to derive this expression from the definition of conditional probability is

$$\begin{aligned} f_{Y|X}(y \mid x) &= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} P(Y \leq y \mid x \leq X \leq x + \varepsilon) \\ &= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \frac{P(\{Y \leq y\} \cap \{x \leq X \leq x + \varepsilon\})}{P(x \leq X \leq x + \varepsilon)} \\ &= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \frac{F(x + \varepsilon, y) - F(x, y)}{F_X(x + \varepsilon) - F_X(x)} \\ &= \frac{\partial}{\partial y} \lim_{\varepsilon \to 0} \frac{\frac{\partial}{\partial x} F(x + \varepsilon, y)}{f_X(x + \varepsilon)} \\ &= \frac{\frac{\partial^2}{\partial x \partial y} F(x, y)}{f_X(x)} \\ &= \frac{f(x, y)}{f_X(x)}. \end{aligned}$$

The **conditional mean** or **conditional expectation** is the function

$$m(x) = E(Y \mid X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x)\, dy.$$

The conditional mean $m(x)$ is a function, meaning that when X equals x, then the expected value of Y is $m(x)$.

Similarly, we define the conditional variance of Y given $X = x$ as

$$\begin{aligned} \sigma^2(x) &= Var(Y \mid X = x) \\ &= E\left((Y - m(x))^2 \mid X = x\right) \\ &= E\left(Y^2 \mid X = x\right) - m(x)^2. \end{aligned}$$

Evaluated at $x = X$, the conditional mean $m(X)$ and conditional variance $\sigma^2(x)$ are random variables, functions of X. We write this as $E(Y \mid X) = m(X)$ and $Var(Y \mid X) = \sigma^2(X)$. For example, if $E(Y \mid X = x) = \alpha + \beta x$, then $E(Y \mid X) = \alpha + \beta X$, a transformation of X.

The following are important facts about conditional expectations.

Simple Law of Iterated Expectations:

$$E(E(Y \mid X)) = E(Y) \tag{A.7}$$

Proof:

$$\begin{aligned} E(E(Y \mid X)) &= E(m(X)) \\ &= \int_{-\infty}^{\infty} m(x) f_X(x) dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) f_X(x) dy dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(y, x)\, dy dx \\ &= E(Y). \end{aligned}$$

Law of Iterated Expectations:

$$E(E(Y \mid X, Z) \mid X) = E(Y \mid X) \tag{A.8}$$

Conditioning Theorem. For any function $g(x)$,

$$E(g(X)Y \mid X) = g(X)\, E(Y \mid X) \tag{A.9}$$

Proof: Let

$$\begin{aligned} h(x) &= E(g(X)Y \mid X = x) \\ &= \int_{-\infty}^{\infty} g(x) y f_{Y|X}(y \mid x)\, dy \\ &= g(x) \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x)\, dy \\ &= g(x) m(x) \end{aligned}$$

where $m(x) = E(Y \mid X = x)$. Thus $h(X) = g(X)m(X)$, which is the same as $E(g(X)Y \mid X) = g(X)\, E(Y \mid X)$.

A.7 Transformations

Suppose that $X \in R^k$ with continuous distribution function $F_X(x)$ and density $f_X(x)$. Let $Y = g(X)$ where $g(x) : R^k \to R^k$ is one-to-one, differentiable, and invertible. Let $h(y)$ denote the inverse of $g(x)$. The **Jacobian** is

$$J(y) = \det\left(\frac{\partial}{\partial y'}h(y)\right).$$

Consider the univariate case $k = 1$. If $g(x)$ is an increasing function, then $g(X) \leq Y$ if and only if $X \leq h(Y)$, so the distribution function of Y is

$$\begin{aligned} F_Y(y) &= P\left(g(X) \leq y\right) \\ &= P\left(X \leq h(Y)\right) \\ &= F_X\left(h(Y)\right) \end{aligned}$$

so the density of Y is

$$f_Y(y) = \frac{d}{dy}F_Y(y) = f_X\left(h(Y)\right)\frac{d}{dy}h(y).$$

If $g(x)$ is a decreasing function, then $g(X) \leq Y$ if and only if $X \geq h(Y)$, so

$$\begin{aligned} F_Y(y) &= P\left(g(X) \leq y\right) \\ &= 1 - P\left(X \geq h(Y)\right) \\ &= 1 - F_X\left(h(Y)\right) \end{aligned}$$

and the density of Y is

$$f_Y(y) = -f_X\left(h(Y)\right)\frac{d}{dy}h(y).$$

We can write these two cases jointly as

$$f_Y(y) = f_X\left(h(Y)\right)\left|J(y)\right|. \tag{A.10}$$

This is known as the **change-of-variables** formula. This same formula (A.10) holds for $k > 1$, but its justification requires deeper results from analysis.

As one example, take the case $X \sim U[0,1]$ and $Y = -\ln(X)$. Here, $g(x) = -\ln(x)$ and $h(y) = \exp(-y)$ so the Jacobian is $J(y) = -\exp(y)$. As the range of X is $[0,1]$, that for Y is $[0,\infty)$. Since $f_X(x) = 1$ for $0 \leq x \leq 1$ (A.10) shows that

$$f_Y(y) = \exp(-y), \qquad 0 \leq y \leq \infty,$$

an exponential density.

A.8 Normal and Related Distributions

The **standard normal** density is

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right), \qquad -\infty < x < \infty.$$

This density has all moments finite. Since it is symmetric about zero all odd moments are zero. By iterated integration by parts, we can also show that $EX^2 = 1$ and $EX^4 = 3$. It is conventional to write $X \sim N(0,1)$, and to denote the standard normal density function by $\phi(x)$ and its distribution function by $\Phi(x)$. The latter has no closed-form solution.

If Z is standard normal and $X = \mu + \sigma Z$, then using the change-of-variables formula, X has density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty.$$

which is the **univariate normal density**. The mean and variance of the distribution are μ and σ^2, and it is conventional to write $X \sim N(\mu, \sigma^2)$.

For $x \in R^k$, the **multivariate normal density** is

$$f(x) = \frac{1}{(2\pi)^{k/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{(x-\mu)' \Sigma^{-1} (x-\mu)}{2}\right), \qquad x \in R^k.$$

The mean and covariance matrix of the distribution are μ and Σ, and it is conventional to write $X \sim N(\mu, \Sigma)$.

It useful to observe that the MGF and CF of the multivariate normal are $\exp\left(\lambda'\mu + \lambda'\Sigma\lambda/2\right)$ and $\exp\left(i\lambda'\mu - \lambda'\Sigma\lambda/2\right)$, respectively.

If $X \in R^k$ is multivariate normal and the elements of X are mutually uncorrelated, then $\Sigma = diag\{\sigma_j^2\}$ is a diagonal matrix. In this case the density function can be written as

$$\begin{aligned} f(x) &= \frac{1}{(2\pi)^{k/2} \sigma_1 \cdots \sigma_k} \exp\left(-\left(\frac{(x_1-\mu_1)^2/\sigma_1^2 + \cdots + (x_k-\mu_k)^2/\sigma_k^2}{2}\right)\right) \\ &= \prod_{j=1}^{k} \frac{1}{(2\pi)^{1/2} \sigma_j} \exp\left(-\frac{(x_j-\mu_j)^2}{2\sigma_j^2}\right) \end{aligned}$$

which is the product of marginal univariate normal densities. This shows that if X is multivariate normal with uncorrelated elements, then they are mutually independent.

Another useful fact is that if $X \sim N(\mu, \Sigma)$ and $Y = a + BX$ with B an invertible matrix, then by the change-of-variables formula, the density of Y is

$$f(y) = \frac{1}{(2\pi)^{k/2} \det(\Sigma_Y)^{1/2}} \exp\left(-\frac{(y-\mu_Y)' \Sigma_Y^{-1} (y-\mu_Y)}{2}\right), \qquad x \in R^k.$$

where $\mu_Y = a + B\mu$ and $\Sigma_Y = B\Sigma B'$, where we used the fact that $\det(B\Sigma B')^{1/2} = \det(\Sigma)^{1/2} \det(B)$. This shows that linear transformations of normals are also normal.

Theorem A.8.1 *Let $X \sim N(0, I_r)$ and set $Q = X'X$. Q has the density*

$$f(y) = \frac{1}{\Gamma\left(\frac{r}{2}\right) 2^{r/2}} y^{r/2-1} \exp(-y/2), \qquad y \geq 0. \tag{A.11}$$

*and is known as the **chi-square** density with r degrees of freedom, denoted χ_r^2. Its mean and variance are $\mu = r$ and $\sigma^2 = 2r$.*

Theorem A.8.2 *If $Z \sim N(0, A)$ with $A > 0$, $q \times q$, then $Z'A^{-1}Z \sim \chi_q^2$.*

Theorem A.8.3 *Let $Z \sim N(0,1)$ and $Q \sim \chi^2_r$ be independent. Set*

$$t_r = \frac{Z}{\sqrt{Q/r}}.$$

The density of t_r is

$$f(x) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{\pi r}\Gamma\left(\frac{r}{2}\right)\left(1+\frac{x^2}{r}\right)^{\frac{r+1}{2}}} \tag{A.12}$$

*and is known as the **student's t distribution** with r degrees of freedom.*

Proof of Theorem A.8.1. The MGF for the density (A.11) is

$$\begin{aligned} E\exp(tQ) &= \int_0^\infty \frac{1}{\Gamma\left(\frac{r}{2}\right)2^{r/2}} y^{r/2-1}\exp(ty)\exp(-y/2)\,dy \\ &= (1-2t)^{-r/2} \end{aligned} \tag{A.13}$$

where the second equality uses the fact that $\int_0^\infty y^{a-1}\exp(-by)\,dy = b^{-a}\Gamma(a)$, which can be found by applying change-of-variables to the gamma function. For $Z \sim N(0,1)$ the distribution of Z^2 is

$$\begin{aligned} P\left(Z^2 \le y\right) &= 2P\left(0 \le Z \le \sqrt{y}\right) \\ &= 2\int_0^{\sqrt{y}} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)dx \\ &= \int_0^y \frac{1}{\Gamma\left(\frac{1}{2}\right)2^{1/2}} s^{-1/2}\exp\left(-\frac{s}{2}\right)ds \end{aligned}$$

using the change–of-variables $s = x^2$ and the fact $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. Thus the density of Z^2 is (A.11) with $r = 1$. From (A.13), we see that the MGF of Z^2 is $(1-2t)^{-1/2}$. Since we can write $Q = X'X = \sum_{j=1}^r Z_j^2$ where the Z_j are independent $N(0,1)$, (A.6) can be used to show that the MGF of Q is $(1-2t)^{-r/2}$, which we showed in (A.13) is the MGF of the density (A.11). ■

Proof of Theorem A.8.2. The fact that $A > 0$ means that we can write $A = CC'$ where C is non-singular. Then $A^{-1} = C^{-1\prime}C^{-1}$ and

$$C^{-1}Z \sim N(0, C^{-1}AC^{-1\prime}) = N(0, C^{-1}CC'C^{-1\prime}) = N(0, I_q).$$

Thus

$$Z'A^{-1}Z = Z'C^{-1\prime}C^{-1}Z = \left(C^{-1}Z\right)'\left(C^{-1}Z\right) \sim \chi^2_q.$$

■

Proof of Theorem A.8.3. Using the simple law of iterated expectations, t_r has distribution

function

$$\begin{aligned} F(x) &= P\left(\frac{Z}{\sqrt{Q/r}} \leq x\right) \\ &= E\left\{Z \leq x\sqrt{\frac{Q}{r}}\right\} \\ &= E\left[P\left(Z \leq x\sqrt{\frac{Q}{r}} \mid Q\right)\right] \\ &= E\Phi\left(x\sqrt{\frac{Q}{r}}\right) \end{aligned}$$

Thus its density is

$$\begin{aligned} f(x) &= E\frac{d}{dx}\Phi\left(x\sqrt{\frac{Q}{r}}\right) \\ &= E\left(\phi\left(x\sqrt{\frac{Q}{r}}\right)\sqrt{\frac{Q}{r}}\right) \\ &= \int_0^\infty \left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{qx^2}{2r}\right)\right)\sqrt{\frac{q}{r}}\left(\frac{1}{\Gamma\left(\frac{r}{2}\right)2^{r/2}}q^{r/2-1}\exp\left(-q/2\right)\right)dq \\ &= \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)}\left(1+\frac{x^2}{r}\right)^{-\left(\frac{r+1}{2}\right)}. \end{aligned}$$

■

A.9 Maximum Likelihood

If the distribution of Y_i is $F(y,\theta)$ where F is a known distribution function and $\theta \in \Theta$ is an unknown $m \times 1$ vector, we say that the distribution is **parametric** and that θ is the **parameter** of the distribution F. The space Θ is the set of permissible value for θ. In this setting the **method of maximum likelihood** is the appropriate technique for estimation and inference on θ.

If the distribution F is continuous then the density of Y_i can be written as $f(y,\theta)$ and the joint density of a random sample $\tilde{Y} = (Y_1, ..., Y_n)$ is

$$f_n\left(\tilde{Y},\theta\right) = \prod_{i=1}^n f\left(Y_i,\theta\right).$$

The **likelihood** of the sample is this joint density evaluated at the observed sample values, viewed as a function of θ. The **log-likelihood** function is its natural log

$$L_n(\theta) = \sum_{i=1}^n \ln f\left(Y_i,\theta\right).$$

If the distribution F is discrete, the likelihood and log-likelihood are constructed by setting $f(y,\theta) = P(Y = y,\theta)$.

Define the **Hessian**

$$H = -E\frac{\partial^2}{\partial\theta\partial\theta'}\ln f\left(Y_i,\theta_0\right) \tag{A.14}$$

and the **outer product** matrix

$$\Omega = E\left(\frac{\partial}{\partial\theta}\ln f\left(Y_i,\theta_0\right)\frac{\partial}{\partial\theta}\ln f\left(Y_i,\theta_0\right)'\right). \tag{A.15}$$

Two important features of the likelihood are

Theorem A.9.1

$$\frac{\partial}{\partial\theta}E\ln f\left(Y_i,\theta\right)\Big|_{\theta=\theta_0} = 0 \tag{A.16}$$

$$H = \Omega \equiv I_0 \tag{A.17}$$

The matrix I_0 is called the **information**, and the equality (A.17) is often called the **information matrix equality**.

Theorem A.9.2 ***Cramer-Rao Lower Bound.*** *If $\tilde{\theta}$ is an unbiased estimator of $\theta \in R$, then $Var(\tilde{\theta}) \geq (nI_0)^{-1}$.*

The Cramer-Rao Theorem gives a lower bound for estimation. However, the restriction to unbiased estimators means that the theorem has little direct relevance for finite sample efficiency.

The **maximum likelihood estimator** or **MLE** $\hat{\theta}$ is the parameter value which maximizes the likelihood (equivalently, which maximizes the log-likelihood). We can write this as

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, L_n(\theta).$$

In some simple cases, we can find an explicit expression for $\hat{\theta}$ as a function of the data, but these cases are rare. More typically, the MLE $\hat{\theta}$ must be found by numerical methods.

Why do we believe that the MLE $\hat{\theta}$ is estimating the parameter θ? Observe that when standardized, the log-likelihood is a sample average

$$\frac{1}{n}L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\ln f\left(Y_i,\theta\right) \to_p E\ln f\left(Y_i,\theta\right) \equiv L(\theta).$$

As the MLE $\hat{\theta}$ maximizes the left-hand-side, we can see that it is an estimator of the maximizer of the right-hand-side. The first-order condition for the latter problem is

$$0 = \frac{\partial}{\partial\theta}L(\theta) = \frac{\partial}{\partial\theta}E\ln f\left(Y_i,\theta\right)$$

which holds at $\theta = \theta_0$ by (A.16). In fact, under conventional regularity conditions, $\hat{\theta}$ is consistent for this value, $\hat{\theta} \to_p \theta_0$ as $n \to \infty$.

Theorem A.9.3 *Under regularity conditions, $\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to_d N\left(0, I_0^{-1}\right)$.*

Thus in large samples, the approximate variance of the MLE is $(nI_0)^{-1}$ which is the Cramer-Rao lower bound. Thus in large samples the MLE has approximately the best possible variance. Therefore the MLE is called **asymptotically efficient**.

Typically, to estimate the asymptotic variance of the MLE we use an estimate based on the Hessian formula (A.14)

$$\hat{H} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta'} \ln f\left(Y_i, \hat{\theta}\right) \tag{A.18}$$

We then set $\hat{I}_0^{-1} = \hat{H}^{-1}$. Asymptotic standard errors for $\hat{\theta}$ are then the square roots of the diagonal elements of $n^{-1}\hat{I}_0^{-1}$.

Sometimes a parametric density function $f(y,\theta)$ is used to approximate the true unknown density $f(y)$, but it is not literally believed that the model $f(y,\theta)$ is necessarily the true density. In this case, we refer to $L_n(\hat{\theta})$ as a **quasi-likelihood** and the its maximizer $\hat{\theta}$ as a **quasi-mle** or **QMLE.**

In this case there is not a "true" value of the parameter θ. Instead we define the **pseudo-true** value θ_0 as the maximizer of

$$E \ln f\left(Y_i, \theta\right) = \int f\left(y\right) \ln f\left(y, \theta\right) dy$$

which is the same as the minimizer of

$$KLIC = \int f\left(y\right) \ln\left(\frac{f(y)}{f\left(y,\theta\right)}\right) dy$$

the Kullback-Leibler information distance between the true density $f(y)$ and the parametric density $f(y,\theta)$. Thus the QMLE θ_0 is the value which makes the parametric density "closest" to the true value according to this measure of distance. The QMLE is consistent for the pseudo-true value, but has a different covariance matrix than in the pure MLE case, since the information matrix equality (A.17) does not hold. A minor adjustment to Theorem (A.9.3) yields the asymptotic distribution of the QMLE:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to_d N\left(0, V\right), \qquad V = H^{-1}\Omega H^{-1}$$

The moment estimator for V is

$$\hat{V} = \hat{H}^{-1}\hat{\Omega}\hat{H}^{-1}$$

where $\hat{H}$ is given in (A.18) and

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f\left(Y_i, \hat{\theta}\right) \frac{\partial}{\partial\theta} \ln f\left(Y_i, \hat{\theta}\right)'.$$

Asymptotic standard errors (sometimes called qmle standard errors) are then the square roots of the diagonal elements of $n^{-1}\hat{V}$.

Proof of Theorem A.9.1. To see (A.16),

$$\begin{aligned}
\frac{\partial}{\partial\theta}E\ln f\left(Y_i,\theta\right)\Big|_{\theta=\theta_0} &= \frac{\partial}{\partial\theta}\int \ln f\left(y,\theta\right) f\left(y,\theta_0\right) dy\Big|_{\theta=\theta_0} \\
&= \int \frac{\partial}{\partial\theta} f\left(y,\theta\right) \frac{f\left(y,\theta_0\right)}{f\left(y,\theta\right)} dy\Big|_{\theta=\theta_0} \\
&= \frac{\partial}{\partial\theta}\int f\left(y,\theta\right) dy\Big|_{\theta=\theta_0} \\
&= \frac{\partial}{\partial\theta}1\Big|_{\theta=\theta_0} = 0.
\end{aligned}$$

Similarly, we can show that

$$E\left(\frac{\frac{\partial^2}{\partial\theta\partial\theta'}f\left(Y_i,\theta_0\right)}{f\left(Y_i,\theta_0\right)}\right) = 0.$$

By direction computation,

$$\begin{aligned}
\frac{\partial^2}{\partial\theta\partial\theta'}\ln f\left(Y_i,\theta_0\right) &= \frac{\frac{\partial^2}{\partial\theta\partial\theta'}f\left(Y_i,\theta_0\right)}{f\left(Y_i,\theta_0\right)} - \frac{\frac{\partial}{\partial\theta}f\left(Y_i,\theta_0\right)\frac{\partial}{\partial\theta}f\left(Y_i,\theta_0\right)'}{f\left(Y_i,\theta_0\right)^2} \\
&= \frac{\frac{\partial^2}{\partial\theta\partial\theta'}f\left(Y_i,\theta_0\right)}{f\left(Y_i,\theta_0\right)} - \frac{\partial}{\partial\theta}\ln f\left(Y_i,\theta_0\right)\frac{\partial}{\partial\theta}\ln f\left(Y_i,\theta_0\right)'.
\end{aligned}$$

Taking expectations yields (A.17). ■

Proof of Theorem A.9.2.

$$S = \frac{\partial}{\partial\theta}\ln f_n\left(\tilde{Y},\theta_0\right) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f\left(Y_i,\theta_0\right)$$

which by Theorem (A.9.1) has mean zero and variance nH. Write the estimator $\tilde{\theta} = \tilde{\theta}(\tilde{Y})$ as a function of the data. Since $\tilde{\theta}$ is unbiased for any θ,

$$\theta = E\tilde{\theta} = \int \tilde{\theta}(\tilde{y}) f\left(\tilde{y},\theta\right) d\tilde{y}$$

where $\tilde{y} = (y_1, ..., y_n)$. Differentiating with respect to θ and evaluating at θ_0 yields

$$1 = \int \tilde{\theta}(\tilde{y})\frac{\partial}{\partial\theta}f\left(\tilde{y},\theta\right) d\tilde{y} = \int \tilde{\theta}(\tilde{y})\frac{\partial}{\partial\theta}\ln f\left(\tilde{y},\theta\right) f\left(\tilde{y},\theta_0\right) d\tilde{y} = E\left(\tilde{\theta}S\right).$$

By the Cauchy-Schwarz inequality

$$1 = \left|E\left(\tilde{\theta}S\right)\right|^2 \leq Var\left(S\right) Var\left(\tilde{\theta}\right)$$

so

$$Var\left(\tilde{\theta}\right) \geq \frac{1}{Var\left(S\right)} = \frac{1}{nH}.$$

■

Proof of Theorem A.9.3 Taking the first-order condition for maximization of $L_n(\theta)$, and making a first-order Taylor series expansion,

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta} L_n(\theta)\Big|_{\theta=\hat{\theta}} \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f\left(Y_i, \hat{\theta}\right) \\ &\simeq \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f\left(Y_i, \theta_0\right) + \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta'} \ln f\left(Y_i, \theta_n\right)\left(\hat{\theta} - \theta_0\right), \end{aligned}$$

where θ_n lies on a line segment joining $\hat{\theta}$ and θ_0. (Technically, the specific value of θ_n varies by row in this expansion.) Rewriting this equation, we find

$$\left(\hat{\theta} - \theta_0\right) = \left(-\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta'} \ln f\left(Y_i, \theta_n\right)\right)^{-1} \left(\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f\left(Y_i, \theta_0\right)\right).$$

Since $\frac{\partial}{\partial\theta} \ln f\left(Y_i, \theta_0\right)$ is mean-zero with covariance matrix Ω, an application of the CLT yields

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f\left(Y_i, \theta_0\right) \to_d N\left(0, \Omega\right).$$

The analysis of the sample Hessian is somewhat more complicated due to the presence of θ_n. Let $H(\theta) = -\frac{\partial^2}{\partial\theta\partial\theta'} \ln f\left(Y_i, \theta\right)$. If it is continuous in θ, then since $\theta_n \to_p \theta_0$ we find $H(\theta_n) \to_p H$ and so

$$\begin{aligned} -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta'} \ln f\left(Y_i, \theta_n\right) &= \frac{1}{n} \sum_{i=1}^{n} \left(-\frac{\partial^2}{\partial\theta\partial\theta'} \ln f\left(Y_i, \theta_n\right) - H(\theta_n)\right) + H(\theta_n) \\ &\to_p H \end{aligned}$$

by an application of a uniform WLLN. Together,

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to_d H^{-1} N\left(0, \Omega\right) = N\left(0, H^{-1}\Omega H^{-1}\right) = N\left(0, H^{-1}\right),$$

the final equality using Theorem A.9.1 . ■

Appendix B

Numerical Optimization

Many econometric estimators are defined by an optimization problem of the form

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} Q(\theta) \tag{B.1}$$

where the parameter is $\theta \in \Theta \subset R^m$ and the criterion function is $Q(\theta) : \Theta \to R$. For example NLLS, GLS, MLE and GMM estimators take this form. In most cases, $Q(\theta)$ can be computed for given θ, but $\hat{\theta}$ is not available in closed form. In this case, numerical methods are required to obtain $\hat{\theta}$.

B.1 Grid Search

Many optimization problems are either one dimensional ($m = 1$) or involve one-dimensional optimization as a sub-problem (for example, a line search). In this context grid search may be employed.

Grid Search. Let $\Theta = [a, b]$ be an interval. Pick some $\varepsilon > 0$ and set $G = (b - a)/\varepsilon$ to be the number of gridpoints. Construct an equally spaced grid on the region $[a, b]$ with G gridpoints, which is $\{\theta(j) = a + j(b - a)/G : j = 0, ..., G\}$. At each point evaluate the criterion function and find the gridpoint which yields the smallest value of the criterion, which is $\theta(\hat{\jmath})$ where $\hat{\jmath} = \operatorname{argmin}_{0 \le j \le G} Q(\theta(j))$. This value $\theta(\hat{\jmath})$ is the gridpoint estimate of $\hat{\theta}$. If the grid is sufficiently fine to capture small oscillations in $Q(\theta)$, the approximation error is bounded by ε, that is, $\left|\theta(\hat{\jmath}) - \hat{\theta}\right| \le \varepsilon$. Plots of $Q(\theta(j))$ against $\theta(j)$ can help diagnose errors in grid selection. This method is quite robust but potentially costly.

Two-Step Grid Search. The gridsearch method can be refined by a two-step execution. For an error bound of ε pick G so that $G^2 = (b - a)/\varepsilon$ For the first step define an equally spaced grid on the region $[a, b]$ with G gridpoints, which is $\{\theta(j) = a + j(b - a)/G : j = 0, ..., G\}$. At each point evaluate the criterion function and let $\hat{\jmath} = \operatorname{argmin}_{0 \le j \le G} Q(\theta(j))$. For the second step define an equally spaced grid on $[\theta(\hat{\jmath} - 1), \theta(\hat{\jmath} + 1)]$ with G gridpoints, which is $\{\theta'(k) = \theta(\hat{\jmath} - 1) + 2k(b - a)/G^2 : k = 0, ..., G\}$. Let $\hat{k} = \operatorname{argmin}_{0 \le k \le G} Q(\theta'(k))$. The estimate of $\hat{\theta}$ is $\theta'(\hat{k})$. The advantage of the two-step method over a one-step grid search is that the number of function evaluations has been reduced from $(b - a)/\varepsilon$ to $2\sqrt{(b - a)/\varepsilon}$ which can be substantial. The disadvantage is that if the function $Q(\theta)$ is irregular, the first-step grid may not bracket $\hat{\theta}$ which thus would be missed.

B.2 Gradient Methods

Gradient Methods are iterative methods which produce a sequence $\theta_i : i = 1, 2, ...$ which are designed to converge to $\hat{\theta}$. All require the choice of a starting value θ_1, and all require the computation of the **gradient** of $Q(\theta)$

$$g(\theta) = \frac{\partial}{\partial \theta} Q(\theta)$$

and some require the **Hessian**

$$H(\theta) = \frac{\partial^2}{\partial \theta \partial \theta'} Q(\theta).$$

If the functions $g(\theta)$ and $H(\theta)$ are not analytically available, they can be calculated numerically. Take the $j'th$ element of $g(\theta)$. Let δ_j be the $j'th$ unit vector (zeros everywhere except for a one in the $j'th$ row). Then for ε small

$$g_j(\theta) \simeq \frac{Q(\theta + \delta_j \varepsilon) - Q(\theta)}{\varepsilon}.$$

Similarly,

$$g_{jk}(\theta) \simeq \frac{Q(\theta + \delta_j \varepsilon + \delta_k \varepsilon) - Q(\theta + \delta_k \varepsilon) - Q(\theta + \delta_j \varepsilon) + Q(\theta)}{\varepsilon^2}$$

In many cases, numerical derivatives can work well but can be computationally costly relative to analytic derivatives. In some cases, however, numerical derivatives can be quite unstable.

Most gradient methods are a variant of **Newton's method** which is based on a quadratic approximation. By a Taylor's expansion for θ close to $\hat{\theta}$

$$0 = g(\hat{\theta}) \simeq g(\theta) + H(\theta) \left(\hat{\theta} - \theta \right)$$

which implies

$$\hat{\theta} = \theta - H(\theta)^{-1} g(\theta).$$

This suggests the iteration rule

$$\hat{\theta}_{i+1} = \theta_i - H(\theta_i)^{-1} g(\theta_i).$$

where

One problem with Newton's method is that it will send the iterations in the wrong direction if $H(\theta_i)$ is not positive definite. One modification to prevent this possibility is quadratic hill-climbing which sets

$$\hat{\theta}_{i+1} = \theta_i - \left(H(\theta_i) + \alpha_i I_m \right)^{-1} g(\theta_i).$$

where α_i is set just above the smallest eigenvalue of $H(\theta_i)$ if $H(\theta)$ is not positive definite.

Another productive modification is to add a scalar **steplength** λ_i. In this case the iteration rule takes the form

$$\theta_{i+1} = \theta_i - D_i g_i \lambda_i \tag{B.2}$$

where $g_i = g(\theta_i)$ and $D_i = H(\theta_i)^{-1}$ for Newton's method and $D_i = \left(H(\theta_i) + \alpha_i I_m \right)^{-1}$ for quadratic hill-climbing.

Allowing the steplength to be a free parameter allows for a line search, a one-dimensional optimization. To pick λ_i write the criterion function as a function of λ

$$Q(\lambda) = Q(\theta_i + D_i g_i \lambda)$$

a one-dimensional optimization problem. There are two common methods to perform a line search. A **quadratic approximation** evaluates the first and second derivatives of $Q(\lambda)$ with respect to λ, and picks λ_i as the value minimizing this approximation. The **half-step** method considers the sequence $\lambda = 1, 1/2, 1/4, 1/8, \ldots$. Each value in the sequence is considered and the criterion $Q(\theta_i + D_i g_i \lambda)$ evaluated. If the criterion has improved over $Q(\theta_i)$, use this value, otherwise move to the next element in the sequence.

Newton's method does not perform well if $Q(\theta)$ is irregular, and it can be quite computationally costly if $H(\theta)$ is not analytically available. These problems have motivated alternative choices for the weight matrix D_i. These methods are called **Quasi-Newton** methods. Two popular methods are do to Davidson-Fletcher-Powell (DFP) and Broyden-Fletcher-Goldfarb-Shanno (BFGS).

Let

$$\begin{aligned} \Delta g_i &= g_i - g_{i-1} \\ \Delta \theta_i &= \theta_i - \theta_{i-1} \end{aligned}$$

and . The DFP method sets

$$D_i = D_{i-1} + \frac{\Delta\theta_i \Delta\theta_i'}{\Delta\theta_i' \Delta g_i} + \frac{D_{i-1} \Delta g_i \Delta g_i' D_{i-1}}{\Delta g_i' D_{i-1} \Delta g_i}.$$

The BFGS methods sets

$$D_i = D_{i-1} + \frac{\Delta\theta_i \Delta\theta_i'}{\Delta\theta_i' \Delta g_i} - \frac{\Delta\theta_i \Delta\theta_i'}{\left(\Delta\theta_i' \Delta g_i\right)^2} \Delta g_i' D_{i-1} \Delta g_i + \frac{\Delta\theta_i \Delta g_i' D_{i-1}}{\Delta\theta_i' \Delta g_i} + \frac{D_{i-1} \Delta g_i \Delta\theta_i'}{\Delta\theta_i' \Delta g_i}.$$

For any of the gradient methods, the iterations continue until the sequence has converged in some sense. This can be defined by examining whether $|\theta_i - \theta_{i-1}|$, $|Q(\theta_i) - Q(\theta_{i-1})|$ or $|g(\theta_i)|$ has become small.

B.3 Derivative-Free Methods

All gradient methods can be quite poor in locating the global minimum when $Q(\theta)$ has several local minima. Furthermore, the methods are not well defined when $Q(\theta)$ is non-differentiable. In these cases, alternative optimization methods are required. One example is the **simplex method** of Nelder-Mead (1965).

A more recent innovation is the method of **simulated annealing (SA).** For a review see Goffe, Ferrier, and Rodgers (1994). The SA method is a sophisticated random search. Like the gradient methods, it relies on an iterative sequence. At each iteration, a random variable is drawn and added to the current value of the parameter. If the resulting criterion is decreased, this new value is accepted. If the criterion is increased, it may still be accepted depending on the extent of the increase and another randomization. The latter property is needed to keep the algorithm from selecting a local minimum. As the iterations continue, the variance of the random innovations is shrunk. The SA algorithm stops when a large number of iterations is unable to improve the criterion. The SA method has been found to be successful at locating global minima. The downside is that it can take considerable computer time to execute.

Bibliography

[1] Aitken, A.C. (1935): "On least squares and linear combinations of observations," *Proceedings of the Royal Statistical Society*, 55, 42-48.

[2] Akaike, H. (1973): "Information theory and an extension of the maximum likelihood principle." In B. Petroc and F. Csake, eds., *Second International Symposium on Information Theory.*

[3] Anderson, T.W. and H. Rubin (1949): "Estimation of the parameters of a single equation in a complete system of stochastic equations," *The Annals of Mathematical Statistics*, 20, 46-63.

[4] Andrews, D.W.K. (1988): "Laws of large numbers for dependent non-identically distributed random variables,' *Econometric Theory,* 4, 458-467.

[5] Andrews, D.W.K. (1991), "Asymptotic normality of series estimators for nonparameric and semiparametric regression models," *Econometrica*, 59, 307-345.

[6] Andrews, D.W.K. (1993), "Tests for parameter instability and structural change with unknown change point," *Econometrica*, 61, 821-8516.

[7] Andrews, D.W.K. and M. Buchinsky: (2000): "A three-step method for choosing the number of bootstrap replications," *Econometrica*, 68, 23-51.

[8] Andrews, D.W.K. and W. Ploberger (1994): "Optimal tests when a nuisance parameter is present only under the alternative," *Econometrica,* 62, 1383-1414.

[9] Basmann, R. L. (1957): "A generalized classical method of linear estimation of coefficients in a structural equation," *Econometrica,* 25, 77-83.

[10] Bekker, P.A. (1994): "Alternative approximations to the distributions of instrumental variable estimators, *Econometrica,* 62, 657-681.

[11] Billingsley, P. (1968): *Convergence of Probability Measures.* New York: Wiley.

[12] Billingsley, P. (1979): *Probability and Measure.* New York: Wiley.

[13] Bose, A. (1988): "Edgeworth correction by bootstrap in autoregressions," *Annals of Statistics*, 16, 1709-1722.

[14] Breusch, T.S. and A.R. Pagan (1979): "The Lagrange multiplier test and its application to model specification in econometrics," *Review of Economic Studies*, 47, 239-253.

[15] Brown, B.W. and W.K. Newey (2002): "GMM, efficient bootstrapping, and improved inference ," *Journal of Business and Economic Statistics.*

[16] Carlstein, E. (1986): "The use of subseries methods for estimating the variance of a general statistic from a stationary time series," *Annals of Statistics*, 14, 1171-1179.

[17] Chamberlain, G. (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305-334.

[18] Choi, I. and P.C.B. Phillips (1992): "Asymptotic and finite sample distribution theory for IV estimators and tests in partially identified structural equations," *Journal of Econometrics*, 51, 113-150.

[19] Chow, G.C. (1960): "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, 28, 591-603.

[20] Davidson, J. (1994): *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford: Oxford University Press.

[21] Davison, A.C. and D.V. Hinkley (1997): *Bootstrap Methods and their Application*. Cambridge University Press.

[22] 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.

[23] Donald S.G. and W.K. Newey (2001): "Choosing the number of instruments," *Econometrica*, 69, 1161-1191.

[24] Dufour, J.M. (1997): "Some impossibility theorems in econometrics with applications to structural and dynamic models," *Econometrica*, 65, 1365-1387.

[25] Efron, Bradley (1979): "Bootstrap methods: Another look at the jackknife," *Annals of Statistics*, 7, 1-26.

[26] Efron, Bradley (1982): *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial and Applied Mathematics.

[27] Efron, Bradley and R.J. Tibshirani (1993): *An Introduction to the Bootstrap*, New York: Chapman-Hall.

[28] Eicker, F. (1963): "Asymptotic normality and consistency of the least squares estimators for families of linear regressions," *Annals of Mathematical Statistics*, 34, 447-456.

[29] Engle, R.F. and C.W.J. Granger (1987): "Co-integration and error correction: Representation, estimation and testing," *Econometrica*, 55, 251-276.

[30] Frisch, R. and F. Waugh (1933): "Partial time regressions as compared with individual trends," *Econometrica*, 1, 387-401.

[31] Gallant, A.F. and D.W. Nychka (1987): "Seminonparametric maximum likelihood estimation," *Econometrica*, 55, 363-390.

[32] Gallant, A.R. and H. White (1988): *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. New York: Basil Blackwell.

[33] Goldberger, Arthur S. (1991): *A Course in Econometrics*. Cambridge: Harvard University Press.

[34] Goffe, W.L., G.D. Ferrier and J. Rogers (1994): "Global optimization of statistical functions with simulated annealing," *Journal of Econometrics*, 60, 65-99.

[35] Gauss, K.F. (1809): "Theoria motus corporum coelestium," in *Werke*, Vol. VII, 240-254.

[36] Granger, C.W.J. (1969): "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, 37, 424-438.

[37] Granger, C.W.J. (1981): "Some properties of time series data and their use in econometric specification," *Journal of Econometrics*, 16, 121-130.

[38] Granger, C.W.J. and T. Teräsvirta (1993): *Modelling Nonlinear Economic Relationships*, Oxford University Press, Oxford.

[39] Hall, A. R. (2000): "Covariance matrix estimation and the power of the overidentifying restrictions test," *Econometrica*, 68, 1517-1527,

[40] Hall, P. (1992): *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag.

[41] Hall, P. (1994): "Methodology and theory for the bootstrap," *Handbook of Econometrics, Vol. IV*, eds. R.F. Engle and D.L. McFadden. New York: Elsevier Science.

[42] Hall, P. and J.L. Horowitz (1996): "Bootstrap critical values for tests based on Generalized-Method-of-Moments estimation," *Econometrica*, 64, 891-916.

[43] Hahn, J. (1996): "A note on bootstrapping generalized method of moments estimators," *Econometric Theory*, 12, 187-197.

[44] Hansen, B.E. (1992): "Efficient estimation and testing of cointegrating vectors in the presence of deterministic trends," *Journal of Econometrics*, 53, 87-121.

[45] Hansen, B.E. (1996): "Inference when a nuisance parameter is not identified under the null hypothesis," *Econometrica*, 64, 413-430.

[46] Hansen, L.P. (1982): "Large sample properties of generalized method of moments estimators, *Econometrica*, 50, 1029-1054.

[47] Hansen, L.P., J. Heaton, and A. Yaron (1996): "Finite sample properties of some alternative GMM estimators," *Journal of Business and Economic Statistics*, 14, 262-280.

[48] Hausman, J.A. (1978): "Specification tests in econometrics," *Econometrica*, 46, 1251-1271.

[49] Heckman, J. (1979): "Sample selection bias as a specification error," *Econometrica*, 47, 153-161.

[50] Imbens, G.W. (1997): "One step estimators for over-identified generalized method of moments models," Review of Economic Studies, 64, 359-383.

[51] Imbens, G.W., R.H. Spady and P. Johnson (1998): "Information theoretic approaches to inference in moment condition models," *Econometrica*, 66, 333-357.

[52] Jarque, C.M. and A.K. Bera (1980): "Efficient tests for normality, homoskedasticity and serial independence of regression residuals, *Economic Letters*, 6, 255-259.

[53] Johansen, S. (1988): "Statistical analysis of cointegrating vectors," *Journal of Economic Dynamics and Control*, 12, 231-254.

[54] Johansen, S. (1991): "Estimation and hypothesis testing of cointegration vectors in the presence of linear trend," *Econometrica*, 59, 1551-1580.

[55] Johansen, S. (1995): *Likelihood-Based Inference in Cointegrated Vector Auto-Regressive Models*, Oxford University Press.

[56] Johansen, S. and K. Juselius (1992): "Testing structural hypotheses in a multivariate cointegration analysis of the PPP and the UIP for the UK," *Journal of Econometrics*, 53, 211-244.

[57] Kitamura, Y. (2001): "Asymptotic optimality and empirical likelihood for testing moment restrictions," *Econometrica*, 69, 1661-1672.

[58] Kitamura, Y. and M. Stutzer (1997): "An information-theoretic alternative to generalized method of moments," *Econometrica*, 65, 861-874..

[59] Koenker, Roger (2005): *Quantile Regression*. Cambridge University Press.

[60] Kunsch, H.R. (1989): "The jackknife and the bootstrap for general stationary observations," *Annals of Statistics*, 17, 1217-1241.

[61] Kwiatkowski, D., P.C.B. Phillips, P. Schmidt, and Y. Shin (1992): "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?" *Journal of Econometrics*, 54, 159-178.

[62] Lafontaine, F. and K.J. White (1986): "Obtaining any Wald statistic you want," *Economics Letters*, 21, 35-40.

[63] Lovell, M.C. (1963): "Seasonal adjustment of economic time series," *Journal of the American Statistical Association*, 58, 993-1010.

[64] MacKinnon, J.G. (1990): "Critical values for cointegration," in Engle, R.F. and C.W. Granger (eds.) *Long-Run Economic Relationships: Readings in Cointegration*, Oxford, Oxford University Press.

[65] MacKinnon, J.G. and H. White (1985): "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 29, 305-325.

[66] Magnus, J. R., and H. Neudecker (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, New York: John Wiley and Sons.

[67] Muirhead, R.J. (1982): *Aspects of Multivariate Statistical Theory*. New York: Wiley.

[68] Nelder, J. and R. Mead (1965): "A simplex method for function minimization," *Computer Journal*, 7, 308-313.

[69] Newey, W.K. and K.D. West (1987): "Hypothesis testing with efficient method of moments estimation," *International Economic Review*, 28, 777-787.

[70] Owen, Art B. (1988): "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, 75, 237-249.

[71] Owen, Art B. (2001): *Empirical Likelihood*. New York: Chapman & Hall.

[72] Phillips, P.C.B. (1989): "Partially identified econometric models," *Econometric Theory*, 5, 181-240.

[73] Phillips, P.C.B. and S. Ouliaris (1990): "Asymptotic properties of residual based tests for cointegration," *Econometrica*, 58, 165-193.

[74] Politis, D.N. and J.P. Romano (1996): "The stationary bootstrap," *Journal of the American Statistical Association*, 89, 1303-1313.

[75] Potscher, B.M. (1991): "Effects of model selection on inference," *Econometric Theory*, 7, 163-185.

[76] Qin, J. and J. Lawless (1994): "Empirical likelihood and general estimating equations," *The Annals of Statistics*, 22, 300-325.

[77] Ramsey, J. B. (1969): "Tests for specification errors in classical linear least-squares regression analysis," *Journal of the Royal Statistical Society*, Series B, 31, 350-371.

[78] Rudin, W. (1987): *Real and Complex Analysis*, 3rd edition. New York: McGraw-Hill.

[79] Said, S.E. and D.A. Dickey (1984): "Testing for unit roots in autoregressive-moving average models of unknown order," *Biometrika*, 71, 599-608.

[80] Shao, J. and D. Tu (1995): *The Jackknife and Bootstrap*. NY: Springer.

[81] Sargan, J.D. (1958): "The estimation of economic relationships using instrumental variables," *Econometrica, 2*6, 393-415.

[82] Sheather, S.J. and M.C. Jones (1991): "A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society, Series B*, 53, 683-690.

[83] Shin, Y. (1994): "A residual-based test of the null of cointegration against the alternative of no cointegration," *Econometric Theory*, 10, 91-115.

[84] Silverman, B.W. (1986): *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

[85] Sims, C.A. (1972): "Money, income and causality," *American Economic Review*, 62, 540-552.

[86] Sims, C.A. (1980): "Macroeconomics and reality," *Econometrica, 48*, 1-48.

[87] Staiger, D. and J.H. Stock (1997): "Instrumental variables regression with weak instruments," *Econometrica,* 65, 557-586.

[88] Stock, J.H. (1987): "Asymptotic properties of least squares estimators of cointegrating vectors," *Econometrica,* 55, 1035-1056.

[89] Stock, J.H. (1991): "Confidence intervals for the largest autoregressive root in U.S. macroeconomic time series," *Journal of Monetary Economics,* 28, 435-460.

[90] Stock, J.H. and J.H. Wright (2000): "GMM with weak identification," *Econometrica,* 68, 1055-1096.

[91] Theil, H. (1953): "Repeated least squares applied to complete equation systems," The Hague, Central Planning Bureau, mimeo.

[92] Tobin, J. (1958): "Estimation of relationships for limited dependent variables," *Econometrica, 2*6, 24-36.

[93] Wald, A. (1943): "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Transactions of the American Mathematical Society*, 54, 426-482.

[94] Wang, J. and E. Zivot (1998): "Inference on structural parameters in instrumental variables regression with weak instruments," *Econometrica,* 66, 1389-1404.

[95] White, H. (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 48, 817-838.

[96] White, H. (1984): *Asymptotic Theory for Econometricians,* Academic Press.

[97] 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.