Econometrics with *gretl*

Proceedings of the

gretl Conference 2009

Bilbao, Spain, May 28-29

I. Díaz-Emparanza, P. Mariel, M.V. Esteban
(Editors)

eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Econometrics with gretl

Proceedings of the *gretl Conference 2009*
Bilbao, Spain, May 28-29, 2009

Econometrics with gretl

Proceedings of the

gretl Conference 2009

Bilbao, Spain, May 28-29, 2009

I. Díaz-Emparanza, P. Mariel, M.V. Esteban
(Editors)

Editors:

Ignacio Díaz-Emparanza
Departamento de Economía Aplicada III (Econometría y Estadística)
Facultad de Ciencias Económicas y Empresariales
Universidad del País Vasco
Avenida Lehendakari Aguirre, 83
48015 Bilbao
Spain

Petr Mariel
Departamento de Economía Aplicada III (Econometría y Estadística)
Facultad de Ciencias Económicas y Empresariales
Universidad del País Vasco
Avenida Lehendakari Aguirre, 83
48015 Bilbao
Spain

María Victoria Esteban
Departamento de Economía Aplicada III (Econometría y Estadística)
Facultad de Ciencias Económicas y Empresariales
Universidad del País Vasco
Avenida Lehendakari Aguirre, 83
48015 Bilbao
Spain

ISBN: 978-84-692-2600-1
Depósito Legal: BI-1428-09

Preface

This proceedings volume contains the papers presented at the GRETL CONFERENCE held in Bilbao, Spain, May 28-29, 2009.

The *gretl* project (GNU Regression, Econometrics and Time Series Library) was initially promoted at the beginning of 2000 by Allin Cottrell, who took, as the basis for it, the econometric program "ESL", originally developed by Ramu Ramanathan and published under an *open source* license. It was around August 2000 when the Free Software Foundation decided to accept gretl as a GNU program. It was at that point when the community of gretl users started to grow. In 2004 Riccardo (Jack) Lucchetti started his collaboration with the project, acting as a second programmer.

The use of the `gettext` program has really been the turning point for gretl to become a more international software. In fact, gretl is currently translated to Basque (Susan Orbe and Marian Zubia), Czech (the CERGE-EI team lead by J. Hanousek), German (Markus Hahn and Sven Schreiber), French (Florent Bresson and Michel Robitaille), Italian (Cristian Rigamonti), Polish (Pawel Kufel, Tadeusz Kufel and Marcin Blazejowski), Portuguese (Hélio Guilherme), Brazilian Portuguese (Hélio Guilherme and Henrique Andrade), Russian (Alexander B. Gedranovich), Spanish (Ignacio Díaz-Emparanza), Turkish (A. Talha Yalta), and, pretty soon, Chinese (Y. N. Yang).

The 2009 gretl Conference has been the first gretl program users and developers meeting. The conference has been organized as a small scientific meeting, having three invited conferences given by Allin Cottrell, Stephen Pollock and Jack Lucchetti, to whom conference organizers wish to give their special thanks for their collaboration and willingness to participate in the conference. Papers presented in the different sessions scheduled at the conference have contributed to develop and extend our knowledge about gretl and some other free software programs for econometric computations and teaching. Allin Cottrell talked about the evolution and development of gretl from its beginnings and, in addition, he also described which ones would be, in his view, the future challenges and paths this interesting project could follow. Stephen Pollock introduced IDEOLOG, a program that allows the filtering of economic time series and that also includes several novel specific filtering procedures that mainly operate in the frequency domain. Jack Lucchetti presented an econometric analysis of the data from the gretl downloads from SourceForge. His findings suggested that, even though gretl has become a fundamental tool for teaching Economet-

rics, it has not yet been widely accepted as a computation program for research in Economics.

Organizers wish to thank specially authors who sent their papers for evaluation. Their active collaboration has allowed organizers to be able to have four very interesting and differently focused sessions in the scientific programme for this conference. The first one dealt with *Econometric Theory*, the second one centered on *Econometric Applications* , the third one on the use of *Free-software programs for teaching Econometrics*, and the last one concentrated on *Contributions to the development of gretl*.

The editors of this volume wish to show their most sincere appreciation and thanks to Josu Arteche, Giorgio Calzolari, Michael Creel, Josef Jablonsky, and Tadeusz Kufel, for their effort and dedication to the conference success as members of the Scientific Committee. We also wish to thank specially those authors presenting papers appearing in this proceedings volume, for their careful manuscript preparation so that our editing work was a lot easier to handle. Finally, we wish to thank Vicente Núñez-Antón for his valuable help in the preparation and organization of this conference.

Ignacio, Petr and Maria Victoria

Bilbao, April 2009

Organization

The gretl Conference 2009 is organized by the Department of Applied Economics III (Econometrics and Statistics), the University of the Basque Country and the School of Economics and Business Administration in cooperation with the *gretl Development Team*.

Organizing Commitee

Conference Chair:	Ignacio Díaz-Emparanza (University of the Basque Country, Spain)
Co-Chair:	Petr Mariel (University of the Basque Country, Spain)
Members:	Maria Victoria Esteban (University of the Basque Country, Spain)
	Allin Cottrell (Wake Forest University, USA)
	Ricardo (Jack) Lucchetti (Università Politecnica delle Marche, Italy)
	A. Talha Yalta (TOBB University of Economics and Technology, Turkey)

Scientific Committee

Josu Arteche	(University of the Basque Country, Spain)
Giorgio Calzolari	(University of Firenze, Italy)
Michael Creel	(Universitat Autònoma de Barcelona, Spain)
Josef Jablonsky	(University of Economics, Prague, Czech Republic)
Tadeusz Kufel	(Nicolaus Copernicus University, Poland)

Sponsoring Institutions

Department of Education of the Basque Government through grant IT-334-07 (UPV/EHU Econometrics Research Group)
EUSTAT (Basque Institute of Statistics)
Bilbao Turismo and Convention Bureau

Table of Contents

Teaching Econometrics with Free Software

Contributions to gretl Development

Invited Lectures

Gretl: Retrospect, Design and Prospect

Allin Cottrell

Department of Economics, Wake Forest University.
cottrell@wfu.edu

Abstract. In this paper I will give a brief overview of the history of gretl's development, comment on some issues relating to gretl's overall design, and set out some thoughts on gretl's future.

1 Retrospect

Gretl's "modern history" began in September, 2001, when the gretl code was first imported to CVS at `sourceforge.net`. The program's "pre-history" goes back to the mid-1990s. I will briefly describe this background.

1.1 Econometrics software on DOS and Windows 3.0

I began teaching econometrics at Wake Forest University in 1989. I had taught a related course, Business Statistics, at Elon College in North Carolina over the previous few years, and had tried using various statistical packages—RATS and PcGive in particular. Both of these were powerful programs but neither was particularly user-friendly for undergraduates with little computing background. "EZ-RATS" was a brave attempt at user-friendliness but not a great success: it regular crashed and lost my students' work. When I started at Wake Forest I tried a different approach, using Ramanathan's (1989) textbook, which came with its own DOS software, Ecslib (later known as ESL). Ecslib offered a limited range of estimators—OLS, TSLS, and a few FGLS variants—but it was stable, free (as in beer, to users of the textbook) and easy to learn.

Over the first half of the 1990s Microsoft Windows 3.0 (released in May, 1990) became increasingly popular and DOS software came to look dated and unfriendly. I decided to learn to program in Microsoft's Visual Basic. This was not a pretty language, but it did make for easy construction of a Graphical User Interface (GUI). Also in the early '90s I was introduced to TEX and LATEX by a computer-scientist friend, and I used Visual Basic to write GUI "front-ends" for both Ramanathan's ESL and what was at the time the best free implementation of TEX for the PC, Eberhard Mattes' emTEX.[1]

[1] Archive item: the web page for my emTEX front-end is still viewable at `http://www.wfu.edu/economics/ftp/emtexgi.html`.

Since it will be of some relevance in the sequel, let me define a "front-end". I mean a GUI program whose *raison d'être* is to make it easier for a user to interact with a command-driven (or CLI, "Command line interface") program, that is, a program that is driven either by commands typed at an interactive prompt or by a "script" of commands previously written to file. A front-end supplies an apparatus of dialog boxes, buttons, drop-down lists and so on, by means of which it enables the user to formulate a request to the CLI program. The front end then

1. translates the user's request into commands intelligible to the CLI program;
2. feeds these commands to the CLI program;
3. retrieves the output from the CLI program; and
4. displays the output to the user in a "window" of some sort.

From the user's point of view the attraction of such a front-end is that it obviates the need to master the command-line vocabulary and syntax of the underlying CLI program. From the programmer's point of view, it can be an interesting intellectual challenge to take one's knowledge of the CLI program and parlay it into an easy-to-use interface. This offers the same sort of reward as teaching: the satisfaction of taking something difficult and making it as clear and simple as possible.

Anyway, gretl's first precursor was ESLWIN, a Windows front-end for Ramanathan's DOS program ESL.

1.2 Linux comes on the scene

In the mid-1990s Wake Forest University set out an ambitious *Plan for the Class of 2000* which would put it among the most "wired" universities in the US. This involved distributing IBM ThinkPads to all students and faculty,[2] and a team of IBM people came to campus to discuss the plan. Windows 3.11 had reached the end of the road and Windows 95 was about to appear, but IBM's OS/2 could still (just about) be presented as a credible alternative, and the IBM guys gave out copies of OS/2 to faculty members who were willing to give it a try. I tried it but didn't like it much. But in carrying out the experiment with OS/2 I found that it wasn't really all that hard to install a parallel operating system, and that made me think of installing Linux, about which I had been hearing good things.

Linux was very much to my liking from the start, and I have used it almost exclusively since 1995. Since I didn't want to "waste time" using any OS other than Linux, but was still using Ramanathan's ESL with my students, I asked Ramu if he'd be willing to give me a copy of the source code for ESL so that

[2] This was long before IBM sold its ThinkPad business to Lenovo.

I could build a Linux version. He kindly said Yes. Around this time I had a sabbatical semester and spent much of the time learning the C programming language. ESL was written in C, and it didn't take much effort to get it running on Linux.

My first attempt at a GUI econometrics program on Linux was a re-write of ESLWIN using the GUI scripting language Tcl/Tk, namely TkESL: again, a "front-end" for a command-line program. This was workable but I soon felt the need for something better. Front-ends are inherently limited. The external relation between the GUI apparatus and the underlying command-processor is a problem—there's always the possibility of a disconnect when translating from GUI objects to commands, then translating back from text output to GUI display. The smallest change in the CLI program can wreak havoc. Besides, the mechanism is inherently inefficient: too much parsing and re-parsing is required,[3] and for each command, or batch of commands, passed to the CLI program, that program must be run from scratch, which always involves costs of initialization. All of the burden of "remembering the state" is placed on the GUI wrapper.

1.3 Enter GTK

The graphical image manipulation program GIMP—initially written by Spencer Kimball and Peter Mattis when they were graduate students at Berkeley—was one of the first "modern" open-source GUI programs to emerge, and it quickly became a flagship product for the Free Software movement.[4] Mattis had originally used Motif—a proprietary graphical interface toolkit for unix-type systems—for GIMP, but by the time of the 0.6x series he was "really fed up with Motif" and decided to write his own toolkits, which he called gtk and gdk for the Gimp Tool Kit and the Gimp Drawing Kit. By version 0.99 of GIMP (1997) this had evolved into GTK+, and as the historian on `www.gimp.org` relates, "Some developers got the crazy idea that it was a great toolkit and should be used in everything." GTK+ became the basis for the Gnome desktop, and (in a smaller way) it also became the basis for the gretl GUI.

Previously available GUI toolkits for unix/Linux were proprietary and/or very "low-level" and difficult to program. GTK+ introduced a new paradigm

[3] I have no expertise in biology, but I enjoy reading popular science. Over the years I've often been somewhat puzzled by accounts of various sorts of "transcription" of information at sub-cellular level: isn't there more transcription going on than is strictly required? But having programmed GUI front-ends I think I now understand what's happening. Evolution is a hack—a brilliant hack, but a hack nonetheless—and as such it partakes of the same sort of hackery as a GUI front-end, where information that is "well understood" at point A cannot be communicated directly to point B, but must be coded up for re-parsing at B!

[4] The first public release of GIMP was version 0.54 (January 1996). See `http://www.gimp.org/about/ancient_history.html`.

and it was quickly apparent that this was the wave of the future. In the late '90s I first experimented by coding gstar, a GTK+ front end for the "starchart" program (written by Alan Paeth and Craig Counterman),[5] and then began making a proper econometrics GUI using GTK+.

2 Design

Ramanathan's ESL was an all-in-one command-line program.[6] If gretl was to be more than a front-end for that program, the first task was to take the basic econometric code and put it into the form of a library, preferably a "shared" one of the modern sort. The next step was to reconstitute a working command-line program linked against the library, and check that it produced the same results as the original ESL. And the step after that would be to write a GUI client program for the same library—not just an external "front end" but an integrated program.

I'll spare you the details of this process, and just note that it was a great learning experience. I remember my excitement when gretlcli plus libgretl first churned out a set of OLS estimates that checked out correctly against ESL.

The GUI took longer, of course, but eventually it fell into place too. When I started with GTK+ it was clearly a good way to go for the Linux platform, but I began to wonder if I'd have to learn Windows programming if I ever wanted to produce a similar GUI for Microsoft Windows (e.g. for my students). Fortunately this was not so. Thanks to Tor Lillqvist's efforts in porting GTK+ to Windows (originally because he wanted to port the GIMP), programmers working on Linux can now create Windows versions of their GTK+ programs with ease.

2.1 Design schema

Gretl's design schema is shown in Figure 1. The command-line and GUI clients are in a sense at par as clients of libgretl (although of course the GUI program is a great deal more complicated). There are two other elements in the picture.

First, we have implemented several of the less commonly used features in gretl as dynamically loadable modules ("plugins"). The basic idea here was to avoid "bloat" in libgretl itself and hold down the memory footprint of the main library. This may now seem over-scrupulous given the amount of RAM to be found in today's PCs. You may blame my thrifty Scottish upbringing if you wish.

[5] See `http://ricardo.ecn.wfu.edu/~cottrell/gstar/`.

[6] Although the 'L' in ESL stood for "Library", ESL did not provide a library in the technical sense.

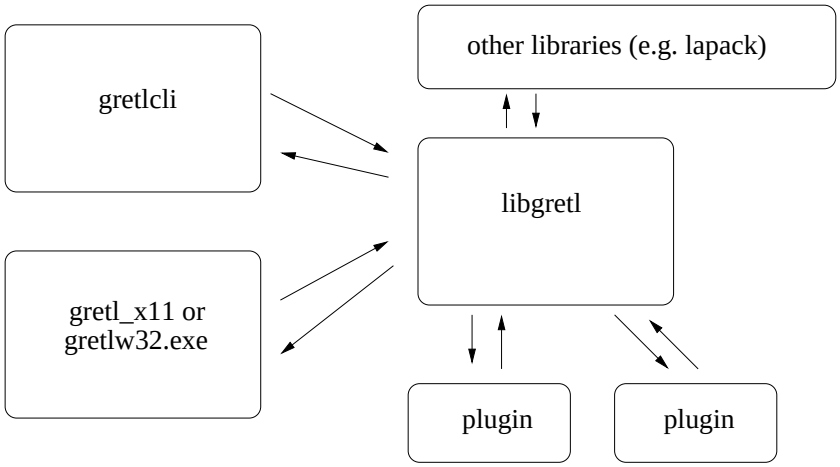


Fig. 1. Schema of gretl's design

Second, gretl relies on various third-party libraries to support its functionality. Probably the most basic of these are libxml2 (since gretl's data files and session files are stored in XML) and Lapack.

Lapack was introduced as a gretl dependency in version 1.0.8 of gretl. This followed the review of gretl by Baiocchi and Distaso (2003). These authors were basically enthusiastic about gretl, but drew attention to some issues of numerical precision. Up till this point all regression calculations in gretl had used the Cholesky decomposition code inherited from ESL. Cholesky decomposition is efficient and accurate for "reasonable" data, but breaks down on very highly collinear data. To ensure accurate results for a data matrix arbitrarily close to singularity it was clear that we needed QR and/or SVD methods, and rather than attempt to code these from scratch we decided to use what is in effect the gold standard for linear computation, Lapack.

I've heard that some people have expressed surprise that gretl doesn't make use of GSL, the Gnu Scientific Library. When I was first thinking about drawing on a third-party library for numerical computation I did consider GSL, but at that time it seemed to me that GSL was relatively immature; it appeared to replicate some but not all of Lapack's functionality, and was much less well tested than the latter. Things have moved on since then, and there may be a case for revisiting this issue, but I can't say I'm sorry to have chosen Lapack. The Fortran interface is awkward at first but it doesn't take long to learn, and we now have a substantial suite of `gretl_matrix` functions that offer C wrappers for the Lapack APIs. And Lapack development continues, witness the recent release of Lapack 3.2.1 and the xblas library with extended precision BLAS functions.

2.2 Other third-party libraries

It's not necessary to enumerate all the additional libraries that gretl either requires or uses optionally (from libfftw to gtksourceview), but a few issues may be worth mentioning.

One issue is Internet connectivity. For several years we've had a "database server" at Wake Forest University from which gretl users can download database files. More recently we've extended this to traffic in "function package" files, and this month (May 2009) I've added the ability to download from within gretl the packages of data files associated with the textbooks by Wooldridge, Stock and Watson, Verbeek and so on. The code we use to enable such traffic was originally "borrowed" from GNU wget and modified for gretl. It seems to work OK for the most part, but I wonder if we could do better, in terms of robustness and extensibility, by linking against libcurl (see `http://curl.haxx.se/`). One nice thing about linking to a third-party library that is under active development is that one gets bug-fixes and new features "for free"—just sit back and enjoy.

Another issue relates to the reading and writing of files in the PKZIP format. Gretl sessions files are zipped in this format, following the pattern of ODF files; and we now read ODS spreadsheets. Similarly to the borrowing from wget, the gretl code for handling zip archives is adapted from code by Mark Adler *et al* from Info-ZIP (zip version 2.31). Since I made that adaptation, libgsf—the Gnome Structured File library, coded in conjunction with Gnumeric—has become reasonably mature. It still (as of version 1.14.12) does not offer all the functionality of Info-ZIP for handling zipfiles, but while our chunk of zip code is effectively frozen it's likely that libgsf will continue to develop apace, so there may be a case for switching to libgsf at some point.

The third and last point that I'll raise in this context does not concern a library, but a third-party *program* of which we make extensive use, namely gnuplot. In this case gretl plays the role of a "front-end", and I spoke earlier of the inherent limitations of that role. There has been talk in the gnuplot community at various times of making a "libgnuplot" library, which would be very helpful from our point of view, but there doesn't seem to be much momentum behind that move, so far as I can see. Over the years I have from time to time checked out some seemingly promising options for graphing libraries, but have not found anything that offers all the functionality of gnuplot. Meanwhile, although there's no gnuplot library yet, gnuplot—which has been very full-featured for a long time—continues to improve. And we've established good relations with the gnuplot developers, and have gained acceptance for some patches which make gnuplot more "gretl-friendly".

This is one area where things are easier in relation to the gretl packages for Windows and OS X than for gretl on Linux. With the Windows and OS X packages we can ship a build of CVS gnuplot that we know will "do the right thing"—specifically, that it can handle UTF-8 encoding, and will produce high-quality PNG and PDF output using the Pango and Cairo libraries. On Linux, gnuplot is probably already installed, but we can't be sure what version and how well it's configured, so we have to implement a lot of workarounds. There may be a case for making a comprehensive gretl package for Linux that bundles gnuplot, though this rather goes against the grain.

3 Prospect

Gretl began life as a teaching tool—specifically, a tool for teaching econometrics at the undergraduate level—but it has grown far beyond that.

As an index of this growth, let me refer back to 2005, when I visited gretl contributors in Ancona, Torun and Bilbao. I recall a discussion in Bilbao where people were putting forward their wish lists for future developments: these included general purpose MLE functionality, GMM, and a general facility for manipulating matrices. I'll admit that these tasks seemed quite daunting at the time, but there followed a flurry of gretl activity in which all of these things were added and more. After spending time together in person, Jack Lucchetti and I were able to collaborate very effectively in coding some of the more challenging additions. The `mle` command was introduced in gretl 1.5.0 (December 2005) and general matrix functionality in 1.5.1 (March 2006). Version 1.6.0 (September 2006) introduced "native" exact ML estimation of ARMA models using the Kalman filter. GMM followed in 1.6.1 (February 2007), along with the Arellano–Bond dynamic panel estimator.

Over the same period the internationalization of gretl accelerated. The first translations were into French and Spanish (2002), then Italian and Polish (2004). Since then we have added Basque, German, Turkish, Russian, Portuguese, Traditional Chinese and most recently Czech. And these are challenging translations, requiring a firm grasp of technical econometric terminology.

The question arises, what should be gretl's role? What should we be aiming for? Reverting to the discussions of 2005 for a moment, there was some debate at the time on the gretl mailing list as to whether it really made sense for gretl to aim for a substantially higher level of econometric sophistication—the alternative being to concentrate on polishing gretl as a robust and user-friendly teaching tool. De facto, this debate has been resolved in favour of the pursuit of sophistication. I think there's at least one good rationale for this.

It has occasionally been put to me that maybe I'm not doing my students a service by teaching econometrics using gretl. The argument is that students are better served if they use from the start a software package that they can continue to use as researchers and professionals; by doing so they would be learning marketable skills and avoiding the need to re-learn how to do things with "standard" software. This sort of comment rankled greatly, but I recognize it has some validity. If gretl were a "dead end" it would be relatively difficult to justify its use in teaching, even if it is more user-friendly than the alternatives. Why not use Stata, Eviews or SAS? Just because we're Free Software ideologues, or happen to enjoy messing about with coding?

3.1 Promoting the adoption of gretl

For several reasons, those of us involved in gretl's development would like to see gretl used more widely, both in teaching and in research. We believe that free, open-source software is desirable in its own right, and is particularly desirable in the scientific domain where it ought to be clear precisely how results are obtained (which is not the case with closed-source proprietary software). [Add reference to Talha Yalta's paper?] We also believe that we have an excellent piece of software in gretl and that students would benefit from using it.[7] And we'd like to see the gretl developer community expand, so that we are less reliant on just a few coders and the project can become self-sustaining.[8]

Jack Lucchetti's analysis of downloads of gretl from the SourceForge site (Lucchetti, 2009) shows a rising trend. That's good, but there are factors making it difficult for gretl to achieve a "break through" to a substantially higher level of adoption. Jack mentions some of these; I'll elaborate a little.

One obvious point is that gretl is competing in a tough market. I don't have solid data to back up this claim, but it seems that Stata and Eviews are currently the leading products in the teaching of econometrics. These programs are also widely used in research, and seem to have edged previously popular software such as RATS and Limdep into niche roles. If gretl were a commercial product aiming to break into this market in a big way (and not just to find a niche) we'd be spending a great deal on advertising, would have a booth at the annual meetings of the Allied Social Science Association, and so on. In fact, of course, from the start we had to rely on "osmosis", achieved through, for example, personal contacts and web searches (e.g. people looking specifically for open-source statistical software). However, we now have a factor working in

[7] This opinion is not confined to gretl developers. I have received many, many emails over the years from professors and students around the world, to just this effect.

[8] And, of course, at a personal level, a bit more recognition would not go amiss: those of use who work on gretl make no money out of it—that was not the plan—but we're only human.

our favour, namely the gretl-aware textbooks that have been published in Polish (Kufel, 2007, also available in a Russian edition) and Spanish (Gallastegui, 2005). In addition we have Lee Adkins' gretl-based ebook (2009) and the forthcoming fourth edition of Christopher Dougherty's *Introduction to Econometrics* (Oxford). In this context I wonder if it would be worth exploring the possibility of writing collaboratively an English-language econometrics text that makes use of gretl? Lee Adkins has done a great deal in this direction already, but I'm thinking of something that would not be tied to a specific existing text (Adkins' ebook is designed to accompany Hill, Griffiths and Lim (2008)), and that, hopefully, could be placed with a major publisher.

3.2 Gretl and R

Still under the general topic of the difficulty of breaking into the highly competitive market in econometric software, one special issue arises for gretl. Besides its specific design features, gretl's most notable attribute is obviously that it is open source and free. But gretl is not entering an empty space in that respect: in residence is the highly respected and full-featured GNU R.

It's noteworthy that R achieved a very positive write-up in the *New York Times* earlier this year (Vance, 2009). Hal Varian (now chief economist at Google) is quoted as saying, "The great beauty of R is that you can modify it to do all sorts of things. And you have a lot of prepackaged stuff that's already available, so you're standing on the shoulders of giants." The article notes the increasing adoption of R for data analysis in both academic and commercial contexts and cites Max Kuhn, associate director of nonclinical statistics at Pfizer: "R has really become the second language for people coming out of grad school now, and there's an amazing amount of code being written for it. You can look on the SAS message boards and see there is a proportional downturn in traffic." The rejoinder by a SAS spokesperson, "We have customers who build engines for aircraft. I am happy they are not using freeware when I get on a jet," sounds defensive and out of touch.

What does R's success mean for gretl? On the one hand it demonstrates that there is scope for free software to make substantial inroads on the turf of the vendors of proprietary statistical software, which is very encouraging. On the other hand it may be seen as raising the question of whether gretl is really needed. Is there room for gretl alongside R in this domain? The gretl developers are well aware of R's strengths but consider that gretl still has a role to play. Gretl has an intuitive GUI; R does not. While gretl is in general much less comprehensive than R it nonetheless adds some specialized econometric functionality in relation to R. And we have taken pains to make gretl as interoperable with R as possible, so that users can take advantage of the complementarity between the

two programs. Nonetheless, this is an area where more thought and more work are required: what exactly should be the relationship between gretl and R?

3.3 Extending gretl

I mentioned earlier the goal that gretl should become self-sustaining, and not overly dependent on the work of a few individuals. In that regard, the way in which the range of packages for R has mushroomed (see Varian's comment above) is very pertinent. In 2006 we introduced a facility to create (and download) "function packages" containing user-contributed code for gretl—code written in the gretl scripting language rather than C. I think this is the right way to go, but although some excellent packages have been contributed it's fair to say that this has not "taken off" to date. This is something we should revisit. There are some awkward aspects of the gretl function packager and we should resolve these. We also need to think about the issue more generally: is there anything we can do specifically to promote the contribution of packages? What can we learn from R?

In closing I'll mention one other aspect of extending gretl and getting it to be better known. In section 2 I spoke about gretl's shared library, libgretl, which was originally created by adapting Ramanathan's ESL code base. The thing is that libgretl in some ways still bears the marks of its origins. Basically, it contains *all* the common code that is needed by both the command-line and the GUI client programs. Some of this code (in particular sections that have been added relatively recently) is quite general and offers a reasonably clean and consistent API—for example the `gretl_matrix` code, the probability distribution code based on Stephen Moshier's `cephes`, the BFGS maximizer, the Kalman filter—while some of it is highly gretl-specific and presents APIs that are unlikely to be comprehensible to anyone who hasn't worked on gretl for years.

One idea for the future then, is to factor out the "private" and "public" components of the current libgretl and to spruce up the APIs of the latter. We'd have, say, `libgretl_priv` and (public) libgretl. It would become easier for new contributors to find their way around the code base, and at the same time third-party developers would have access to a cleaner and more manageable econometrics library for use in their own projects, hence promoting the use of gretl code and gretl's visibility.

Bibliography

[1] ADKINS, L. (2009): *Using gretl for Principles of Econometrics, 3rd edition*. online., Version 1.211, http://www.learneconometrics.com/gretl/ebook.pdf

[2] BAIOCCHI, G. AND DISTASO, W. (2003): "GRETL: Econometric software for the GNU generation", *Journal of Applied Econometrics*, 18, pp. 105–10.

[3] GALLASTEGUI, A. F. (2005): *Econometrìa*, Madrid: Pearson Educación.

[4] HILL, R. C., GRIFFITHS, W. E. AND G. C. LIM (2008) *Principles of Econometrics*, 3e, New York: Wiley.

[5] KUFEL, TADEUSZ (2007) *Ekonometria*, 2e, Warsaw: Wydawnictwo Naukowe PWN.

[6] LUCCHETTI, RICCARDO (JACK) (2009), "Who uses gretl? An analysis of the SourceForge download data", this volume.

[7] RAMANATHAN, RAMU (1989), *Introductory Econometrics with Applications*, San Diego: Harcourt Brace Jovanovich.

[8] VANCE, ASHLEE (2009) "Data Analysts Captivated by R's Power", The New York Times, January 6 http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html Visited 2009-04-04.

IDEOLOG: A Program for Filtering Econometric Data—A Synopsis of Alternative Methods

D.S.G. Pollock

University of Leicester
email: d.s.g.pollock@le.ac.uk

Abstract. An account is given of various filtering procedures that have been implemented in a computer program, which can be used in analysing econometric time series. The program provides some new filtering procedures that operate primarily in the frequency domain. Their advantage is that they are able to achieve clear separations of components of the data that reside in adjacent frequency bands in a way that the conventional time-domain methods cannot.

Several procedures that operate exclusively within the time domain have also been implemented in the program. Amongst these are the bandpass filters of Baxter and King and of Christiano and Fitzgerald, which have been used in estimating business cycles. The Henderson filter, the Butterworth filter and the Leser or Hodrick–Prescott filter are also implemented. These are also described in this paper

Econometric filtering procedures must be able to cope with the trends that are typical of economic time series. If a trended data sequence has been reduced to stationarity by differencing prior to its filtering, then the filtered sequence will need to be re-inflated. This can be achieved within the time domain via the summation operator, which is the inverse of the difference operator. The effects of the differencing can also be reversed within the frequency domain by recourse to the frequency-response function of the summation operator.

1 Introduction

This paper gives an account of some of the facilities that are available in a new computer program, which implements various filters that can be used for extracting the components of an economic data sequence and for producing smoothed and seasonally-adjusted data from monthly and quarterly sequences.

The program can be downloaded from the following web address:

`http://www.le.ac.uk/users/dsgp1/`

It is accompanied by a collection of data and by three log files, which record steps that can be taken in processing some typical economic data. Here, we give an account of the theory that lies behind some of the procedures of the program.

The program originated in a desire to compare some new methods with existing procedures that are common in econometric analyses. The outcome has been a comprehensive facility, which will enable a detailed investigation of univariate econometric time series. The program will also serve to reveal the extent to which the results of an economic analysis might be the consequence of the choice of a particular filtering procedure.

The new procedures are based on the Fourier analysis of the data, and they perform their essential operations in the frequency domain as opposed to the time domain. They depend upon a Fourier transform for carrying the data into the frequency domain and upon an inverse transform for carrying the filtered elements back to the time domain. Filtering procedures usually operate exclusively in the time domain. This is notwithstanding fact that, for a proper understanding of the effects of a filter, one must know its frequency-response function.

The sections of this paper give accounts of the various classes of filters that have been implemented in the program. In the first category, to which section 2 is devoted, are the simple finite impulse response (FIR) or linear moving-average filters that endeavour to provide approximations to the so-called ideal frequency-selective filters. Also in this category of FIR filters is the time-honoured filter of Henderson (1916), which is part of a seasonal-adjustment program that is widely used in central statistical agencies.

The second category concerns filters of the infinite impulse response (IIR) variety, which involve an element of feedback. The filters of this category that are implemented in the program are all derived according to the Wiener–Kolmogorov principle. The principle has been enunciated in connection with the filtering of stationary and doubly-infinite data sequences—see Whittle (1983), for example. However, the purpose of the program is to apply these filters to short non stationary sequences. In section 3, the problem of non stationarity is broached, whereas, in section 4, the adaptations that are appropriate to short sequences are explained.

Section 5 deals with the new frequency-domain filtering procedures. The details of their implementation are described and some of their uses are highlighted. In particular, it is shown how these filters can achieve an ideal frequency selection, whereby all of the elements of the data that fall below a given cut-off frequency are preserved and all those that fall above it are eliminated.

2 The FIR filters

One of the purposes in filtering economic data sequences is to obtain a representation of the business cycle that is free from the distractions of seasonal fluctuations and of high-frequency noise. According to Baxter and King (1999),

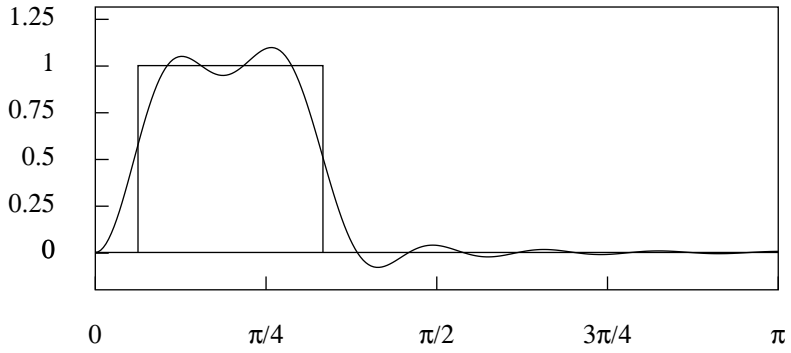


Fig. 1. The frequency response of the truncated bandpass filter of 25 coefficients superimposed upon the ideal frequency response. The lower cut-off point is at $\pi/16$ radians ($11.25°$), corresponding to a period of 32 quarters, and the upper cut-off point is at $\pi/3$ radians ($60°$), corresponding to a period of the 6 quarters.

the business cycle should comprise all elements of the data that have cyclical durations of no less than of one and a half years and not exceeding eight years. For this purpose, they have proposed to use a moving-average bandpass filter to approximate the ideal frequency-selective filter. An alternative approximation, which has the same purpose, has been proposed by Christiano and Fitzgerald (2003). Both of these filters have been implemented in the program.

A stationary data sequence can be resolved into a sum of sinusoidal elements whose frequencies range from zero up to the Nyquist frequency of π radians per sample interval, which represents the highest frequency that is observable in sampled data. A data sequence $\{y_t, t = 0, 1, \ldots, T-1\}$ comprising T observations has the following Fourier decomposition:

$$y_t = \sum_{t-0}^{[T/2]} \{\alpha_j \cos(\omega_j t) + \beta_j \sin(\omega_j t)\}. \tag{1}$$

Here, $[T/2]$ denotes the integer quotient of the division of T by 2. The harmonically related Fourier frequencies $\omega_j = 2\pi j/T; j = 0, \ldots, [T/2]$, which are equally spaced in the interval $[0, \pi]$, are integer multiples of the fundamental frequency $\omega_1 = 2\pi/T$, whereas α_j, β_j are the associated Fourier coefficients, which indicate the amplitudes of the sinusoidal elements of the data sequence. An ideal filter is one that transmits the elements that fall within a specified frequency band, described as the pass band, and which blocks elements at all other frequencies, which constitute the stop band.

In representing the properties of a linear filter, it is common to imagine that it is operating on a doubly-infinite data sequence of a statistically stationary nature. Then, the Fourier decomposition comprises an infinity of sinusoidal elements of negligible amplitudes whose frequencies form a continuum in the interval $[0, \pi]$. The frequency-response function of the filter displays the factors by which the amplitudes of the elements are altered in their passage through the filter.

For an ideal filter, the frequency response is unity within the pass band and zero within the stop band. Such a response is depicted in Figure 1, where the pass band, which runs from $\pi/16$ to $\pi/3$ radians per sample interval, is intended to transmit the elements of a quarterly econometric data sequence that constitute the business cycle.

To achieve an ideal frequency selection with a linear moving-average filter would require an infinite number of filter coefficients. This is clearly impractical; and so the sequence of coefficients must be truncated, wherafter it may be modified in certain ways to diminish the adverse effects of the truncation.

2.1 Approximation to the Ideal Filter

Figure 1 also shows the frequency response of a filter that has been derived by taking twenty-five of the central coefficients of the ideal filter and adjusting their values by equal amounts so that they sum to zero. This is the filter that has been proposed by Baxter and King (1999) for the purpose of extracting the business cycle from economic data. The filter is affected by a considerable leakage, whereby elements that fall within the stop band are transmitted in part by the filter.

The z-transform of a sequence $\{\psi_j\}$ of filter coefficients is the polynomial $\psi(z) = \sum_j \psi_j z$. Constraining the coefficients to sum to zero ensures that the polynomial has a root of unity, which is to say that $\psi(1) = \sum_j \psi_j = 0$. This implies that $\nabla(z) = 1 - z$ is a factor of the polynomial, which indicates that the filter incorporates a differencing operator.

If the filter is symmetric, such that $\psi(z) = \psi_0 + \psi_1(z + z^{-1}) + \cdots + \psi_q(z^q + z^{-q})$ and, therefore, $\psi(z) = \psi(z^{-1})$, then $1 - z^{-1}$ is also a factor. Then, $\psi(z)$ has the combined factor $(1 - z)(1 - z^{-1}) = -z\nabla(z)^2$, which indicates that the filter incorporates a twofold differencing operator. Such a filter is effective in reducing a linear trend to zero; and, therefore, it is applicable to econometric data sequences that have an underlying log-linear trend.

The filter of Baxter and King (1999), which fulfils this condition, is appropriate for the purpose of extracting the business cycle from a trended data sequence. Figure 2 shows the logarithms of data of U.K. real domestic consump-

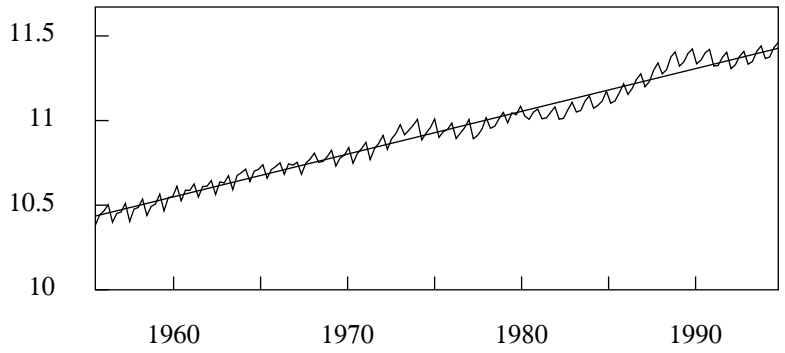


Fig. 2. The quarterly sequence of the logarithms of consumption in the U.K., for the years 1955 to 1994, together with a linear trend interpolated by least-squares regression.

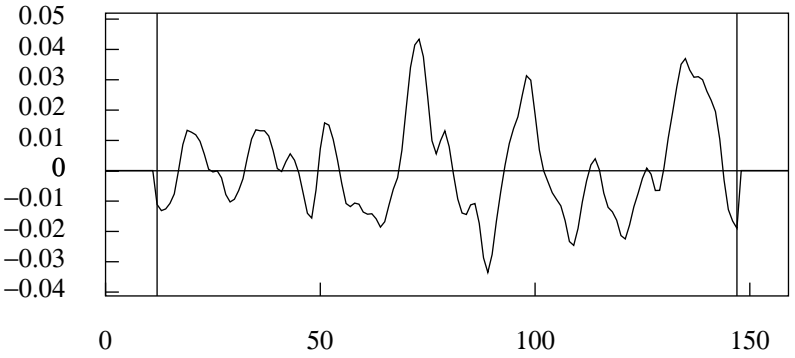


Fig. 3. The sequence derived by applying the truncated bandpass filter of 25 coefficients to the quarterly logarithmic data on U.K. consumption.

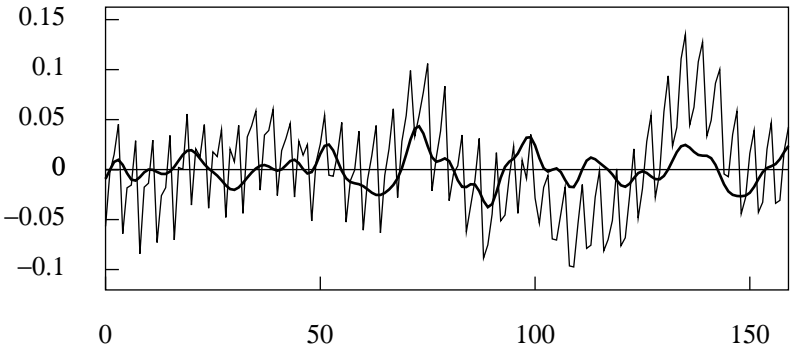


Fig. 4. The sequence derived by applying the bandpass filter of Christiano and Fitzgerald to the quarterly logarithmic data on U.K. consumption.

tion for the years 1955–1994 through which a linear trend has been interpolated. Figure 3 shows the results of subjecting these data to the Baxter–King filter. A disadvantage of the filter, which is apparent in Figure 3, is that it is incapable of reaching the ends of the sample. The first q sample values and the last q remain unprocessed.

To overcome this difficulty, Christiano and Fitzgerald (2003) have proposed a filter with a variable set of coefficients. To generate the filtered value at time t, they associate the central coefficient ψ_0 with y_t. If y_{t-p} falls within the sample, then they associate it with the coefficient ψ_p. Otherwise, if it falls outside the sample, it is disregarded. Likewise, if y_{t+p} falls within the sample, then it is associated with ψ_p, otherwise it is disregarded. If the data follow a first-order random walk, then the first and the last sample elements y_0 and y_{T-1} receive extra weights A and B, which correspond to the sums of the coefficients discarded from the filter at either end. The resulting filtered value at time t may be denoted by

$$\begin{aligned} x_t = Ay_0 + \psi_t y_0 + \cdots + \psi_1 y_{t-1} + \psi_0 y_t \\ + \psi_1 y_{t+1} + \cdots + \psi_{T-1-t} y_{T-1} + By_{T-1}. \end{aligned} \tag{2}$$

This equation comprises the entire data sequence $y_0, \ldots, y_{T-1}$; and the value of t determines which of the coefficients of the infinite-sample filter are involved in producing the current output. The value of x_0 is generated by looking forwards to the end of the sample, whereas the value of x_{T-1} is generated by looking backwards to the beginning of the sample.

For data that appear to have been generated by a first-order random walk with a constant drift, it is appropriate to extract a linear trend before filtering the residual sequence. Figure 4 provides an example of the practice. In fact, this has proved to be the usual practice in most circumstances.

Within the category of FIR filters, the program also implements the time honoured smoothing filter of Henderson (1916), which forms an essential part of the detrending procedure of the X-11 program of the Bureau of the Census. This program provides the method of seasonal adjustment that is used predominantly by central statistical agencies.

Here, the end-of-sample problem is overcome by supplementing the Henderson filter with a set of asymmetric filters that can be applied to the elements of the first and the final segments. These are the Musgrave (1964) filters. (See Quenneville, Ladiray and Lefranc, 2003 for a recent account of these filters.) In the X-11 ARIMA variant, which is used by Statistics Canada, the alternative recourse is adopted of extrapolating the data beyond the ends of the sample so that it can support a time-invariant filter that does run to the ends.

3 The Wiener–Kolmogorov Filters

The program also provides several filters of the feedback variety that are commonly described as infinite-impulse response (IIR) filters. The filters in question are derived according to the finite-sample Wiener–Kolmogorov principle that has been expounded by Pollock (2000, 2007).

The ordinary theory of Wiener–Kolmogorov filtering assumes a doubly-infinite data sequence $y(t) = \xi(t) + \eta(t) = \{y_t; t = 0, \pm 1, \pm 2, \ldots\}$ generated by a stationary stochastic process. The process is compounded from a signal process $\xi(t)$ and a noise process $\eta(t)$ that are assumed to be statistically independent and to have zero-valued means. Then, the autocovariance generating function of $y(t)$ is given by

$$\gamma_y(z) = \gamma_\xi(z) + \gamma_\eta(z), \tag{3}$$

which is sum of the autocovariance functions of $\xi(t)$ and $\eta(t)$.

The object is to extract estimates of the signal sequence $\xi(t)$ and the noise sequence $\eta(t)$ from the data sequence. The z-transforms of the relevant filters are

$$\beta_\xi(z) = \frac{\gamma_\xi(z)}{\gamma_\xi(z) + \gamma_\eta(z)} = \frac{\psi_\xi(z^{-1})\psi_\xi(z)}{\phi(z^{-1})\phi(z)}, \tag{4}$$

and

$$\beta_\eta(z) = \frac{\gamma_\eta(z)}{\gamma_\xi(z) + \gamma_\eta(z)} = \frac{\psi_\eta(z^{-1})\psi_\eta(z)}{\phi(z^{-1})\phi(z)}. \tag{5}$$

It can been that $\beta_\xi(z) + \beta_\eta(z) = 1$, in view of which the filters can be described as complementary.

The factorisations of the filters that are given on the RHS enable them to be applied via a bi-directional feedback process. In the case of the signal extraction filter $\beta_\xi(z)$, the process in question can be represented by the equations

$$\phi(z)q(z) = \psi_\xi(z)y(z) \qquad \text{and} \qquad \phi(z^{-1})x(z) = \psi_\xi(z^{-1})q(z^{-1}), \tag{6}$$

wherein $q(z)$, $y(z)$ and $x(z)$ stand for the z-transforms of the corresponding sequences $q(t)$, $y(t)$ and $x(t)$.

To elucidate these equations, we may note that, in the first of them, the expression associated with z^t is

$$\sum_{j=0}^{m} \phi_j q_{t-j} = \sum_{j=0}^{n} \psi_{\xi,j} y_{t-j}. \tag{7}$$

Given that $\phi_0 = 1$, this serves to determine the value of q_t. Moreover, given that the recursion is assumed to be stable, there need be no restriction on the range

of t. The first equation, which runs forward in time, generates an intermediate output $q(t)$. The second equation, which runs backwards in time, generates the final filtered output $x(t)$.

3.1 Filters for Trended Data

The classical Wiener–Kolmogorov theory can be extended in a straightforward way to cater for non stationary data generated by integrated autoregressive moving-average (ARIMA) processes in which the autoregressive polynomial contains roots of unit value. Such data processes can be described by the equation

$$y(z) = \frac{\delta(z)}{\nabla^p(z)} + \eta(z) \quad \text{or, equivalently,} \quad \nabla^p(z)y(z) = \delta(z) + \nabla^p(z)\eta(z), \tag{8}$$

where $\delta(z)$ and $\eta(z)$ are, respectively, the z-transforms of the mutually independent stationary stochastic sequences $\delta(t)$ and $\eta(t)$, and where $\nabla^p(z) = (1-z)^p$ is the p-th power of the difference operator.

Here, there has to be some restriction on the range of t together with the condition that the elements δ_t and η_t are finite within this range. Also, the z-transforms must comprise the appropriate initial conditions, which are effectively concealed by the notation. (See Pollock 2008 on this point.)

Within the program, two such filters have been implemented. The first is the filter of Leser (1961) and of Hodrick and Prescott (1980, 1997), which is designed to extract the non stationary signal or trend component when the data are generated according to the equation

$$\nabla^2(z)y(z) = g(z) = \delta(z) + \nabla^2(z)\eta(z), \tag{9}$$

where $\delta(t)$ are $\eta(t)$ are mutually independent sequences of independently and identically distributed random variables, generated by so-called white-noise processes. With $\gamma_\delta(z) = \sigma_\delta^2$ and $\gamma_\xi(z) = \sigma_\delta^2\nabla(z^{-1})\nabla(z)$ and with $\gamma_\eta(z) = \sigma_\eta^2$, the z-transforms of the relevant filters become

$$\beta_\xi(z) = \frac{1}{1 + \lambda\nabla^2(z^{-1})\nabla^2(z)}, \tag{10}$$

and

$$\beta_\eta(z) = \frac{\nabla^2(z^{-1})\nabla^2(z)}{\lambda^{-1} + \nabla^2(z^{-1})\nabla^2(z)}, \tag{11}$$

where $\lambda = \sigma_\eta^2/\sigma_\delta^2$, which is described as the smoothing parameter.

The frequency-response functions of the filters for various values of λ are shown in Figure 5. These are obtained by setting $z = e^{-\mathrm{i}\omega} = \cos(\omega) - \mathrm{i}\sin(\omega)$

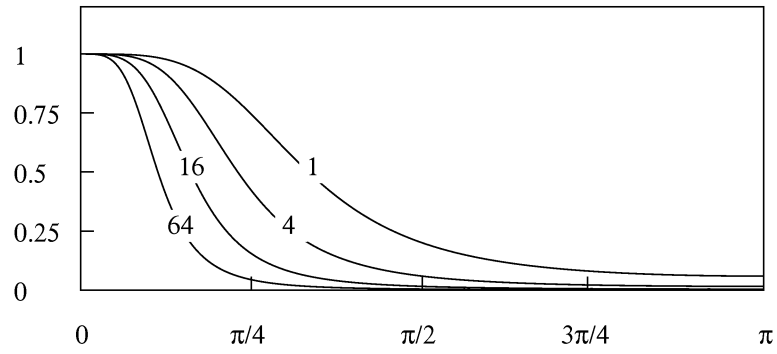


Fig. 5. The frequency-response function of the Hodrick–Prescott smoothing filter for various values of the smoothing parameter λ.

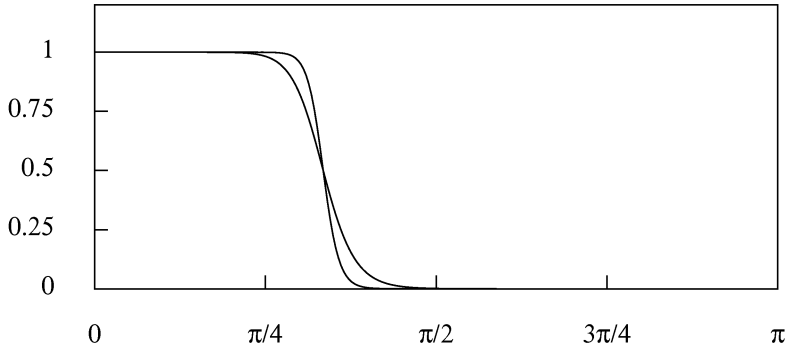


Fig. 6. The frequency-response function of the lowpass Butterworth filters of orders $n = 6$ and $n = 12$ with a nominal cut-off point of $2\pi/3$ radians.

in the formula of (10) and by letting ω run from 0 to π. (In the process, the imaginary quantities are cancelled so as to give rise to the real-valued functions that are plotted in the diagram.)

It is notable that the specification of the underlying process $y(t)$, in which both the signal component $\xi(z) = \delta(z)/\nabla^2(z)$ and the noise component $\eta(z)$ have spectral density functions that extend over the entire frequency range, precludes the clear separation of the components. This is reflected in the fact that, for all but the highest values λ, the filter transmits significant proportions of the elements at all frequencies.

The second of the Wiener–Kolmogorov filters that are implemented in the program is capable of a much firmer discrimination between the signal and noise than is the Leser (1961) filter. This is the Butterworth (1930) filter, which was

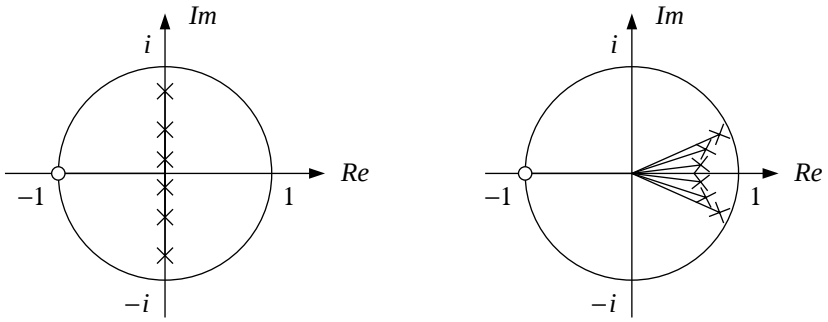


Fig. 7. The pole–zero diagrams of the lowpass Butterworth filters for $n = 6$ when the cut-off is at $\omega = \pi/2$ (left) and at $\omega = \pi/8$.

originally devised as an analogue filter but which can also be rendered in digital form—See Pollock (2000). The filter would be appropriate for extracting the component $(1 + z)^n\delta(z)$ from the sequence

$$g(z) = (1 + z)^n\delta(z) + (1 - z)^n\kappa(z). \tag{12}$$

Here, $\delta(t)$ and $\kappa(t)$ denote independent white-noise processes, whereas there is usually $g(z) = \nabla^2(z)y(z)$, where $y(t)$ is the data process. This corresponds to the case where twofold differencing is required to eliminate a trend from the data. Under these circumstances, the equation of the data process is liable to be represented by

$$\begin{aligned} y(z) &= \xi(z) + \eta(z) \\ &= \frac{(1 + z)^n}{\nabla^2(z)}\delta(z) + (1 - z)^{n-2}\kappa(z). \end{aligned} \tag{13}$$

However, regardless of the degree of differencing to which $y(t)$ must be subjected in reducing it to stationarity, the z-transforms of the complementary filters will be

$$\beta_\xi(z) = \frac{(1 + z^{-1})^n(1 + z)^n}{(1 + z^{-1})^n(1 + z)^n + \lambda(1 - z^{-1})^n(1 - z)^n}, \tag{14}$$

and

$$\beta_\eta(z) = \frac{(1 - z^{-1})^n(1 - z)^n}{\lambda^{-1}(1 + z^{-1})^n(1 + z)^n + (1 - z^{-1})^n(1 - z)^n}, \tag{15}$$

where $\lambda = \sigma_\kappa^2/\sigma_\delta^2$.

It is straightforward to determine the value of λ that will place the cut-off of the filter at a chosen point $\omega_c \in (0, \pi)$. Consider setting $z = \exp\{-\mathrm{i}\omega\}$ in the formula of (14) of the lowpass filter. This gives the following expression for the gain:

$$\begin{aligned}\beta_\xi(e^{-\mathrm{i}\omega}) &= \frac{1}{1 + \lambda\left(\mathrm{i}\dfrac{1 - e^{-\mathrm{i}\omega}}{1 + e^{-\mathrm{i}\omega}}\right)^{2n}} \\ &= \frac{1}{1 + \lambda\big\{\tan(\omega/2)\big\}^{2n}}.\end{aligned} \tag{16}$$

At the cut-off point, the gain must equal $1/2$, whence solving the equation $\beta_\xi(\exp\{-\mathrm{i}\omega_c\}) = 1/2$ gives $\lambda = \{1/\tan(\omega_c/2)\}^{2n}$.

Figure 6 shows how the rate of the transition of the Butterworth frequency response between the pass band and the stop band is affected by the order of the filter. Figure 7 shows the pole–zero diagrams of filters with different cut-off points. As the cut-off frequency is reduced, the transition between the two bands becomes more rapid. Also, some of the poles of the filter move towards the perimeter of the unit circle.

3.2 A Filter for Seasonal Adjustment

The Wiener–Kolmogorov principle is also used in deriving a filter for the seasonal adjustment of monthly and quarterly econometric data. The filter is derived from a model that combines a white-noise component $\eta(t)$ with a seasonal component obtained by passing an independent white noise $\nu(t)$ through a rational filter with poles located on the unit circle at angles corresponding to the seasonal frequencies and with corresponding zeros at the same angles but located inside the circle. The z-transform of the output sequence gives

$$\begin{aligned}y(z) &= \eta(z) + \frac{R(z)}{S(z)}\nu(z) \quad \text{or} \\ S(z)y(z) &= S(z)\eta(z) + R(z)\nu(z),\end{aligned} \tag{17}$$

where

$$R(z) = 1 + \rho z + \rho^2 z^2 + \cdots + \rho^{s-1} z^{s-1} \tag{18}$$

with $\rho < 1$, and

$$S(z) = 1 + z + z^2 + \cdots + z^{s-1}. \tag{19}$$

The z-transform of the seasonal-adjustment filter is

$$\beta(z) = \frac{\sigma_\eta^2 S(z)S(z^{-1})}{S(z)S(z^{-1})\sigma_\eta^2 + \sigma_\nu^2 R(z)R(z^{-1})}. \tag{20}$$

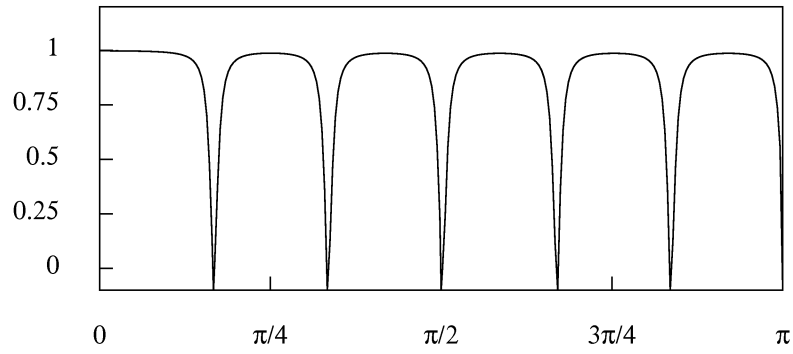


Fig. 8. The gain of a filter for the seasonal adjustment of monthly data. The defining parameters are $\rho = 0.9$ and $\lambda = \sigma_\eta^2/\sigma_\nu^2 = 0.125$.

Setting $z = \exp\{-\mathrm{i}\omega\}$ and letting ω run from 0 to π generates the frequency response of the filter, of which the modulus or gain is plotted in Figure 8 for the case where $\rho = 0.9$ and $\lambda = \sigma_\eta^2/\sigma_\nu^2 = 0.125$.

4 The Finite-Sample Realisations of the W–K Filters

To derive the finite-sample version of a Wiener–Kolmogorov filter, we may consider a data vector $y = [y_0, y_1, \ldots, y_{t-1},]'$ that has a signal component ξ and a noise component η:

$$y = \xi + \eta. \tag{21}$$

The two components are assumed to be independently normally distributed with zero means and with positive-definite dispersion matrices. Then,

$$\begin{aligned} E(\xi) &= 0, \qquad D(\xi) = \Omega_\xi, \\ E(\eta) &= 0, \qquad D(\eta) = \Omega_\eta, \\ \text{and} \quad C(\xi, \eta) &= 0. \end{aligned} \tag{22}$$

A consequence of the independence of ξ and η is that $D(y) = \Omega_\xi + \Omega_\eta$.

The estimates of ξ and η, which may be denoted by x and h respectively, are derived according to the following criterion:

$$\text{Minimise} \quad S(\xi, \eta) = \xi' \Omega_\xi^{-1} \xi + \eta' \Omega_\eta^{-1} \eta \quad \text{subject to} \quad \xi + \eta = y. \tag{23}$$

Since $S(\xi, \eta)$ is the exponent of the normal joint density function $N(\xi, \eta)$, the resulting estimates may be described, alternatively, as the minimum chi-square estimates or as the maximum-likelihood estimates.

Substituting for $\eta = y - \xi$ gives the concentrated criterion function $S(\xi) = \xi'\Omega_\xi^{-1}\xi + (y-\xi)'\Omega^{-1}(y-\xi)$. Differentiating this function in respect of ξ and setting the result to zero gives a condition for a minimum, which specifies the estimate x. This is $\Omega_\eta^{-1}(y-x) = \Omega_\xi^{-1}x$, which, on pre multiplication by Ω_η, can be written as $y = x - \Omega_\eta\Omega_\xi^{-1}x = (\Omega_\xi + \Omega_\eta)\Omega_\xi^{-1}x$. Therefore, the solution for x is

$$x = \Omega_\xi(\Omega_\xi + \Omega_\eta)^{-1}y. \tag{24}$$

Moreover, since the roles of ξ and η are interchangeable in this exercise, and, since $h + x = y$, there are also

$$h = \Omega_\eta(\Omega_\xi + \Omega_\eta)^{-1}y \qquad \text{and} \qquad x = y - \Omega_\eta(\Omega_\xi + \Omega_\eta)^{-1}y. \tag{25}$$

The filter matrices $B_\xi = \Omega_\xi(\Omega_\xi + \Omega_\eta)^{-1}$ and $B_\eta = \Omega_\eta(\Omega_\xi + \Omega_\eta)^{-1}$ of (24) and (25) are the matrix analogues of the z-transforms displayed in equations (4) and (5).

A simple procedure for calculating the estimates x and h begins by solving the equation

$$(\Omega_\xi + \Omega_\eta)b = y \tag{26}$$

for the value of b. Thereafter, one can generate

$$x = \Omega_\xi b \qquad \text{and} \qquad h = \Omega_\eta b. \tag{27}$$

If Ω_ξ and Ω_η correspond to the narrow-band dispersion matrices of moving-average processes, then the solution to equation (26) may be found via a Cholesky factorisation that sets $\Omega_\xi + \Omega_\eta = GG'$, where G is a lower-triangular matrix with a limited number of nonzero bands. The system $GG'b = y$ may be cast in the form of $Gp = y$ and solved for p. Then, $G'b = p$ can be solved for b. The procedure has been described by Pollock (2000).

4.1 Filters for Short Trended Sequences

To adapt these estimates to the case of trended data sequences may require the provision of carefully determined initial conditions with which to start the recursive processes. A variety of procedures are available that are similar, if not identical, in their outcomes. The procedures that are followed in the program depend upon reducing the data sequences to stationarity, in one way or another, before subjecting them to the filters. After the data have been filtered, the trend is liable to be restored.

The first method, which is the simplest in concept, requires the trend to be represented by a polynomial function. In some circumstances, when the economy has been experiencing steady growth, the polynomial will serve as a reasonable characterisation of its underlying trajectory. Thus, in the period 1955–1994

a log-linear trend function provides a firm benchmark against which to measure the cyclical fluctuations of the U.K. economy. The residual deviations from this trend may be subjected to a lowpass filter; and the filtered output can be added to the trend to produce a representation of what is commonly described as the trend-cycle component.

It is desirable that the polynomial trend should interpolate the scatter of points at either end of the data sequence. For this purpose, the program provides a method of weighted least-squares polynomial regression with a wide choice of weighting schemes, which allow extra weight to be placed upon the initial and the final runs of observations.

An alternative way of eliminating the trend is to take differences of the data. Usually, twofold differencing is appropriate. The matrix analogue of the second-order backwards difference operator in the case of $T = 5$ is given by

$$\nabla_5^2 = \begin{bmatrix} Q_*' \\ Q' \end{bmatrix} = \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 & 0 \\ \hline 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{array}\right]. \tag{28}$$

The first two rows, which do not produce true differences, are liable to be discarded. In general, the p-fold differences of a data vector of T elements will be obtained by pre multiplying it by a matrix Q' of order $(T - p) \times T$. Applying Q' to equation (21) gives

$$\begin{aligned} Q'y &= Q'\xi + Q'\eta \\ &= \delta + \kappa = g. \end{aligned} \tag{29}$$

The dispersion matrices of the differenced vectors are

$$D(\delta) = \Omega_\delta = Q'D(\Omega_\xi)Q \qquad \text{and} \qquad D(\kappa) = \Omega_\kappa = Q'D(\Omega_\eta)Q. \tag{30}$$

The estimates d and k of the differenced components are given by

$$d = \Omega_\delta(\Omega_\delta + Q'\Omega_\eta Q)^{-1}Q'y \tag{31}$$

and

$$k = Q'\Omega_\eta Q(\Omega_\delta + Q'\Omega_\eta Q)^{-1}Q'y. \tag{32}$$

To obtain estimates of ξ and η, the estimates of their difference versions must be re-inflated via an anti-differencing or summation operator. We begin by ob-

serving that the inverse of ∇_5^2 is a twofold summation operator given by

$$\nabla_5^{-2} = \begin{bmatrix} S_* & S \end{bmatrix} = \left[\begin{array}{cc|ccc} 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 \\ 3 & 2 & 1 & 0 & 0 \\ 4 & 3 & 2 & 1 & 0 \\ 5 & 4 & 3 & 2 & 1 \end{array}\right]. \tag{33}$$

The first two columns, which constitute the matrix S_*, provide a basis for all linear functions defined on $\{t = 0, 1, \ldots, T-1 = 5\}$. The example can be generalised to the case of a p-fold differencing matrix ∇_T^{-p} of order T. However, in the program, the maximum degree of differencing is $p = 2$.

We observe that, if $g_* = Q_*'y$ and $g = Q'y$ are available, then y can be recovered via the equation

$$y = S_*g_* + Sg. \tag{34}$$

In effect, the elements of g_*, which may be regarded as polynomial parameters, provide the initial conditions for the process of summation or integration, which we have been describing as a process of re-inflation.

The equations by which the estimates of ξ and η may be recovered from those of δ and κ are analogous to equation (34). They are

$$x = S_*d_* + Sd \qquad \text{and} \qquad h = S_*k_* + Sk. \tag{35}$$

In this case, the initial conditions d_* and k_* require to be estimated. The appropriate estimates are the values that minimise the function

$$\begin{aligned}(y-x)'\Omega_\eta^{-1}(y-x) &= (y - S_*d_* - Sd)'\Omega_\eta^{-1}(y - S_*d_* - Sd) \\ &= (S_*k_* + Sk)'\Omega_\eta^{-1}(S_*k_* + Sk).\end{aligned} \tag{36}$$

These values are

$$k_* = -(S_*'\Omega_\eta^{-1}S_*)^{-1}S_*'\Omega_\eta^{-1}Sk \tag{37}$$

and

$$d_* = (S_*'\Omega_\eta^{-1}S_*)^{-1}S_*'\Omega_\eta^{-1}(y - Sd). \tag{38}$$

Equations (37) and (38) together with (31) and (32) provide a complete solution to the problem of estimating the components of the data. However, it is possible to eliminate the initial conditions from the system of estimating equations. This can be achieved with the help of the following identity:

$$\begin{aligned}P_* &= S_*(S_*'\Omega_\eta^{-1}S_*)^{-1}S_*'\Omega_\eta^{-1} \\ &= I - \Omega_\eta Q(Q'\Omega_\eta Q)^{-1}Q' = I - P_Q.\end{aligned} \tag{39}$$

In these terms, the equation of (35) for h becomes $h = (I - P_*)Sk = P_Q Sk$. Using the expression for k from (32) together with the identity $Q'S = I_{T-2}$ gives

$$h = \Omega_\eta Q(\Omega_\delta + Q'\Omega_\eta Q)^{-1}Q'y. \tag{40}$$

This can also be obtained from the equation (32) for k by the removal of the leading differencing matrix Q'. It follows immediately that

$$\begin{aligned} x &= y - h \\ &= y - \Omega_\eta Q(\Omega_\delta + Q'\Omega_\eta Q)^{-1}Q'y. \end{aligned} \tag{41}$$

The elimination of the initial conditions is due to the fact that η is a stationary component. Therefore, it requires no initial conditions other than the zeros that are the appropriate estimates of the pre-sample elements. The direct estimate x of ξ does require initial conditions, but, in view of the adding-up conditions of (21), x can be obtained more readily by subtracting from y the estimate h of η, in the manner of equation (41).

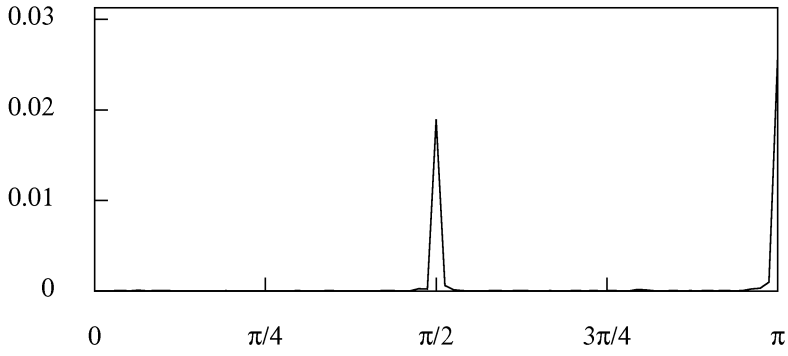


Fig. 9. The periodogram of the first differences of the U.K. logarithmic consumption data.

Observe that, since

$$f = S_*(S_*'S_*)^{-1}S_*'y \tag{42}$$

is an expression for the vector of the ordinates of a polynomial function fitted to the data by an ordinary least-squares regression, the identity of (39) informs us that

$$f = y - Q(Q'Q)^{-1}Q'y \tag{43}$$

is an alternative expression.

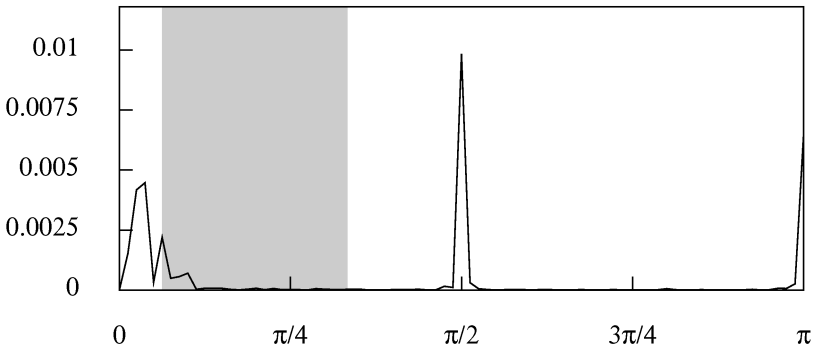


Fig. 10. The periodogram of the residual sequence obtained from the linear detrending of the logarithmic consumption data. A band, with a lower bound of $\pi/16$ radians and an upper bound of $\pi/3$ radians, is masking the periodogram.

The residuals of an OLS polynomial regression of degree p, which are given by $y - f = Q(Q'Q)^{-1}Q'y$, contain same the information as the vector $g = Q'y$ of the p-th differences of the data. The difference operator has the effect of nullifying the element of zero frequency and of attenuating radically the adjacent low-frequency elements. Therefore, the low-frequency spectral structures of the data are not perceptible in the periodogram of the differenced sequence. Figure 9 provides evidence of this.

On the other hand, the periodogram of a trended sequence is liable to be dominated by its low-frequency components, which will mask the other spectral structures. However, the periodogram of the residuals of the polynomial regression can be relied upon to reveal the spectral structures at all frequencies. Moreover, by varying the degree p of the polynomial, one is able to alter the relative emphasis that is given to high-frequency and low-frequency structures. Figure 10 shows that the low-frequency structure of the U.K. consumption data is fully evident in the periodogram of the residuals from fitting a linear trend to the logarithmic data.

4.2 A Flexible Smoothing Filter

A derivation of the estimator of ξ is available that completely circumvents the problem of the initial conditions. This can be illustrated with the case of a generalised version of the Leser (1961) filter in which the smoothing parameter is permitted to vary over the course of the sample. The values of the smoothing parameter are contained in the diagonal matrix $\Lambda = \text{diag}\{\lambda_0, \lambda_1, \ldots, \lambda_{T-1}\}$.

Then, the criterion for finding the vector is to minimise

$$L = (y - \xi)'(y - \xi) + \xi' Q \Lambda Q' \xi. \tag{44}$$

The first term in this expression penalises departures of the resulting curve from the data, whereas the second term imposes a penalty for a lack of smoothness in the curve. The second term comprises $d = Q'\xi$, which is the vector of the p-fold differences of ξ. The matrix Λ serves to generalise the overall measure of the curvature of the function that has the elements of ξ as its sampled ordinates, and it serves to regulate the penalty for roughness, which may vary over the sample.

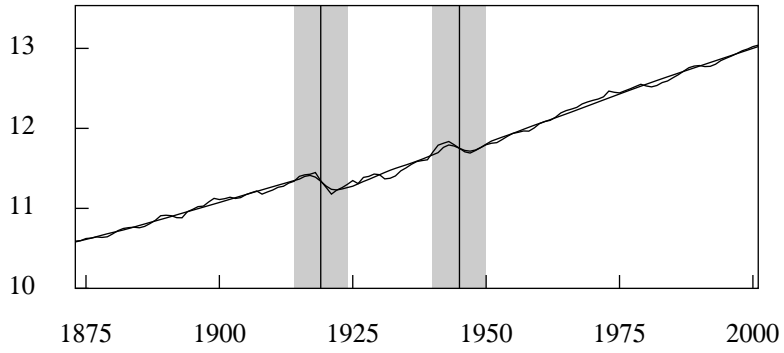


Fig. 11. The logarithms of annual U.K. real GDP from 1873 to 2001 with an interpolated trend. The trend is estimated via a filter with a variable smoothing parameter.

Differentiating L with respect to ξ and setting the result to zero, in accordance with the first-order conditions for a minimum, gives

$$y - x = Q \Lambda Q' x = Q \Lambda d. \tag{45}$$

Multiplying the equation by Q' gives $Q'(y - x) = Q'y - d = Q'Q\Lambda d$, whence $\Lambda d = (\Lambda^{-1} + Q'Q)^{-1} Q'y$. Putting this into the equation $x = y - Q\Lambda d$ gives

$$\begin{aligned} x &= y - Q(\Lambda^{-1} + Q'Q)^{-1} Q'y \\ &= y - \Lambda Q (I + \Lambda Q'Q)^{-1} Q'y. \end{aligned} \tag{46}$$

This filter has been implemented in the program under the guise of a variable smoothing procedure. By giving a high value to the smoothing parameter, a

stiff curve can be generated, which approaches a straight line as $\lambda \to \infty$. On the other hand, structural breaks can be accommodated by greatly reducing the value of the smoothing parameter in their neighbourhood. When $\lambda \to 0$, the filter tends to transmit the unaltered data values.

Figure 11 shown an example of the use of this filter. There were brief disruptions to the steady upwards progress of GDP in the U.K. after the two world wars. These breaks have been absorbed into the trend by reducing the value of the smoothing parameter in their localities. By contrast, the break that is evident in the data following the year 1929 has not been accommodated in the trend.

4.3 A Seasonal-Adjustment Filter

The need for initial conditions cannot be circumvented in cases where the seasonal adjustment filter is applied to trended sequences. Consider the filter that is applied to the differenced data $g = Q'y$ to produce a seasonally-adjusted sequence q. Then, there is

$$q = Q_S(Q_S'Q_S + \lambda^{-1}Q_R'Q_R)^{-1}Q_S'g, \tag{47}$$

where Q_R' and Q_S' are the matrix counterparts of the polynomial operators $R(z)$ and $S(z)$ of (18) and (19) respectively. The seasonally adjusted version of the original trended data will be obtained by re-inflating the filtered sequence q via the equation

$$j = S_*q_* + Sq, \tag{48}$$

where

$$q_* = (S_*'S_*)^{-1}S_*'(y - Sq) \tag{49}$$

is the value that minimises the function

$$(y - j)'(y - j) = (y - S_*q_* + Sq)'(S_*q_* + Sq). \tag{50}$$

5 The Frequency-Domain Filters

Often, in the analysis economic data, we would profit from the availability of a sharp filter, with a rapid transition between the stop band and the pass band that is capable of separating components of the data that lie in closely adjacent frequency bands.

An example of the need for such a filter is provided by a monthly data sequence with an annual seasonal pattern superimposed on a trend–cycle trajectory. The fundamental seasonal frequency is of $\pi/6$ radians or 30 degrees per month, whereas the highest frequency of the trend–cycle component is liable to

exceed $\pi/9$ radians or 20 degrees. This leaves a narrow frequency interval in which a filter that is intended to separate the trend–cycle component from the remaining elements must make the transition from its pass band to its stop band.

To achieve such a sharp transition, an FIR or moving-average filter requires numerous coefficients covering a wide temporal span. Such filters are inappropriate to the short data sequences that are typical of econometric analyses. Rational filters or feedback filters, as we have described them, are capable of somewhat sharper transitions, but they also have their limitations.

When a sharp transition is achieved by virtue of a rational filter with relatively many coefficients, the filter tends to be unstable on account of the proximity of some its poles to the circumference of the unit circle. (See Figure 7 for an example.) Such filters can be excessively influenced by noise contamination in the data and by the enduring effects of ill-chosen initial conditions.

A more effective way of achieving a sharp cut-off is to conduct the filtering operations in the frequency domain. Reference to equation (1) shows that an ideal filter can be obtained by replacing with zeros the Fourier coefficients that are associated with frequencies that fall within the stop band.

5.1 Complex Exponentials and the Fourier Transform

The Fourier coefficients are determined by regressing the data on the trigonometrical functions of the Fourier frequencies according to the following formulae:

$$\alpha_j = \frac{2}{T}\sum_t y_t \cos\omega_j t, \quad \text{and} \quad \beta_j = \frac{2}{T}\sum_t y_t \sin\omega_j t. \tag{51}$$

Also, there is $\alpha_0 = T^{-1}\sum_t y_t = \bar{y}$, and, in the case where $T = 2n$ is an even number, there is $\alpha_n = T^{-1}\sum_t (-1)^t y_t$.

It is more convenient to work with complex Fourier coefficients and with complex exponential functions in place sines and cosines. Therefore, we define

$$\zeta_j = \frac{\alpha_j - \mathrm{i}\beta_j}{2}. \tag{52}$$

Since $\cos(\omega_j t) - \sin(\omega_j t) = e^{-\mathrm{i}\omega_j t}$, it follows that the complex Fourier transform and its inverse are given by

$$\zeta_j = \frac{1}{T}\sum_{t=0}^{T-1} y_t e^{-\mathrm{i}\omega_j t} \quad \longleftrightarrow \quad y_t = \sum_{j=0}^{T-1} \zeta_j e^{\mathrm{i}\omega_j t}, \tag{53}$$

where $\zeta_{T-j} = \zeta_j^* = (\alpha_j + \beta_j)/2$. For a matrix representation of these transforms, one may define

$$U = T^{-1/2}[\exp\{-\mathrm{i}2\pi tj/T\}; t, j = 0, \ldots, T-1], \tag{54}$$
$$\bar{U} = T^{-1/2}[\exp\{\mathrm{i}2\pi tj/T\}; t, j = 0, \ldots, T-1],$$

which are unitary complex matrices such that $U\bar{U} = \bar{U}U = I_T$. Then,

$$\zeta = T^{-1/2}Uy \quad \longleftrightarrow \quad y = T^{1/2}\bar{U}\zeta, \tag{55}$$

where $y = [y_0, y_1, \ldots y_{T-1}]'$ and $\zeta = [\zeta_0, \zeta_1, \ldots \zeta_{T-1}]'$ are the vectors of the data and of their spectral ordinates, respectively.

This notation can be used to advantage for representing the process of applying an ideal frequency-selective filter. Let J be a diagonal selection matrix of order T of zeros and units, wherein the units correspond to the frequencies of the pass band and the zeros to those of the stop band. Then, the selected Fourier ordinates are the nonzero elements of the vector $J\zeta$. By an application of the inverse Fourier transform, the selected elements are carried back to the time domain to form the filtered sequence. Thus, there is

$$x = \bar{U}JUy = \Psi y. \tag{56}$$

Here, $\bar{U}JU = \Psi = [\psi^\circ_{|i-j|}; i, j = 0, \ldots, T-1]$ is a circulant matrix of the filter coefficients that would result from wrapping the infinite sequence of the ideal bandpass coefficients around a circle of circumference T and adding the overlying elements. Thus

$$\psi_k^\circ = \sum_{q=-\infty}^{\infty} \psi_{qT+k}. \tag{57}$$

Applying the wrapped filter to the finite data sequence via a circular convolution is equivalent to applying the original filter to an infinite periodic extension of the data sequence. In practice, the wrapped coefficients of the time-domain filter matrix Ψ would be obtained from the Fourier transform of the vector of the diagonal elements of the matrix J. However, it is more efficient to perform the filtering by operating upon the Fourier ordinates in the frequency domain, which is how the program operates.

The method of frequency-domain filtering can be used to mimic the effects of any linear time-invariant filter, operating in the time domain, that has a well-defined frequency-response function. All that is required is to replace the selection matrix J of equation (56) by a diagonal matrix containing the ordinates of

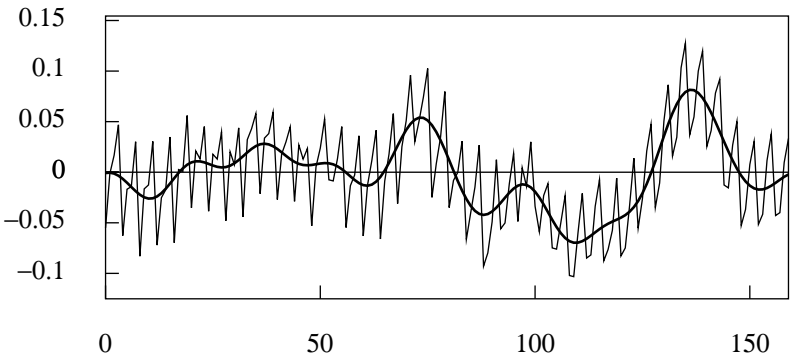


Fig. 12. The residual sequence from fitting a linear trend to the logarithmic consumption data with an interpolated function representing the business cycle.

the desired frequency response, sampled at points corresponding to the Fourier frequencies.

In the case of the Wiener–Kolmogorov filters, defined by equation (24) and (25), one can consider replacing the dispersion matrices Ω_ξ and Ω_η by their circular counterparts

$$\Omega_\xi^\circ = \bar{U}\Lambda_\xi U \qquad \text{and} \qquad \Omega_\eta^\circ = \bar{U}\Lambda_\eta U. \tag{58}$$

Here, Λ_ξ and Λ_η are diagonal matrices containing ordinates sampled from the spectral density functions of the respective processes. The resulting equations for the filtered sequences are

$$x = \Omega_\xi^\circ(\Omega_\xi^\circ + \Omega_\eta^\circ)^{-1}y = \bar{U}\Lambda_\xi(\Lambda_\xi + \Lambda_\eta)^{-1}Uy = \bar{U}J_\xi Uy \tag{59}$$

and

$$h = \Omega_\eta^\circ(\Omega_\xi^\circ + \Omega_\eta^\circ)^{-1}y = \bar{U}\Lambda_\eta(\Lambda_\xi + \Lambda_\eta)^{-1}Uy = \bar{U}J_\eta Uy. \tag{60}$$

An example of the application of the lowpass frequency-domain filter is provided by Figure 12. Here, a filter with a precise cut-off frequency of $\pi/8$ radians has been applied to the residuals from the linear detrending of the logarithms of the U.K. consumption data.

The appropriate cut-off frequency for this filter has been indicated by the periodogram of Figure 10. The smooth curve that has been interpolated through these residuals has been constituted from the Fourier ordinates in the interval $[0, \pi/8]$.

The same residual sequence has also been subjected to the approximate bandpass filter of Christiano and Fitzgerald (2003) to generate the estimated

business cycle of Figure 4. This estimate fails to capture some of the salient low-frequency fluctuations of the data.

The highlighted region Figure 10 also show the extent of the pass band of the bandpass filter; and it appears that the low-frequency structure of the data falls mainly below this band. The fact that, nevertheless, the filter of Christiano and Fitzgerald does reflect a small proportion of the low-frequency fluctuations is due to its substantial leakage over the interval $[0, \pi/16]$, which falls within its nominal stop band.

5.2 Extrapolations and Detrending

To apply the frequency-domain filtering methods, the data must be free of trend. The detrending can be achieved either by differencing the data or by applying the filter to data that are residuals from fitting a polynomial trend. The program has a facility for fitting a polynomial time trend of a degree not exceeding 15. To avoid the problems of collinearity that arise in fitting ordinary polynomials specified in terms of the powers of the temporal index t, a flexible generalised least-squares procedure is provided that depends upon a system of orthogonal polynomials.

In applying the methods, it is also important to ensure that there are no significant disjunctions in the periodic extension of the data at the points where the end of one replication of the sample sequence joins the beginning of the next replication. Equivalently, there must be a smooth transition between the start and finish points when the sequence of T data points is wrapped around a circle of circumference T.

The conventional means of avoiding such disjunctions is to taper the mean-adjusted, detrended data sequence so that both ends decay to zero. (See Bloomfield 1976, for example.) The disadvantage of this recourse is that it falsifies the data at the ends of the sequence, which is particularly inconvenient if, as is often the case in economics, attention is focussed on the most recent data. To avoid this difficulty, the tapering can be applied to some extrapolations, which can be added to the data, either before or after it has been detrended.

In the first case, a polynomial is fitted to the data; and tapered versions of the residual sequence that have been reflected around the endpoints of the sample are added to the extrapolated branches of the polynomial. Alternatively, if the data show strong seasonal fluctuations, then a tapered sequence based on successive repetitions of the ultimate seasonal cycle is added to the upper branch, and a similar sequence based on the first cycle is added to the lower branch.

In the second case, where the data have already been detrended, by the subtraction of a polynomial trend or by the application of the differencing operator, the extrapolations will be added to the horizontal axis.

This method of extrapolation will prevent the end of the sample from being joined directly to its beginning. When the data are supplemented by extrapolations, the circularity of the filter will effect only the furthest points the extrapolations, and the extrapolations will usually be discarded after the filtering has taken place. However, in many cases, extrapolations and their associated tapering will prove to be unnecessary. A case in point is provided by the filtering of the residual sequence of the logarithmic consumption data that is illustrated by Figure 12.

5.3 Anti-Differencing

After a differenced data sequence has been filtered, it will be required to reverse the effects of the differencing via a process of re-inflation. The process can be conducted in the time domain in the manner that has been indicated in section 4, where expressions have been derived for the initial conditions that must accompany the summation operations.

However, if the filtered sequence is the product of a highpass filter and if the original data have been subjected to a twofold differencing operation, then an alternative method of re-inflation is available that operates in the frequency domain. This method is used in the program only if the filtering itself has taken place in the frequency domain.

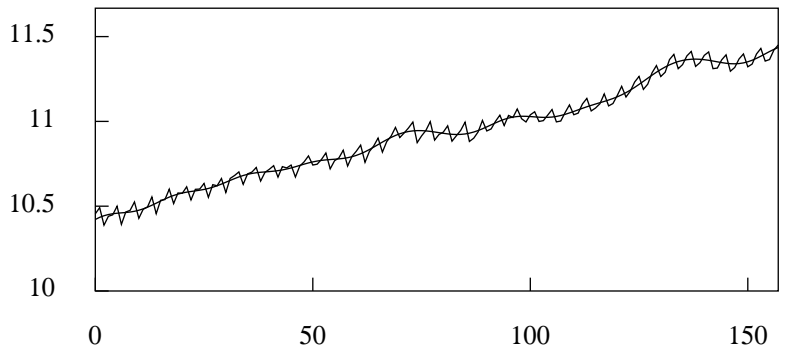


Fig. 13. The trend-cycle component of U.K. consumption determined by the frequency-domain method, superimposed on the logarithmic data.

In that case, the reduction to stationarity will be by virtue of a centralised twofold differencing operator of the form

$$(1 - z^{-1})(1 - z) = -z\nabla^2(z) \tag{61}$$

The frequency-response function of the operator, which is obtained by setting $z = \exp\{-\mathrm{i}\omega\}$ in this equation, is

$$f(\omega) = 2 - 2\cos(\omega). \tag{62}$$

The frequency response of the anti-differencing operator is $v(\omega) = 1/f(\omega)$.

The matrix version of the centralised operator can be illustrated by the case where $T = 5$:

$$N_5 = \begin{bmatrix} n_0' \\ -Q' \\ n_4' \end{bmatrix} = - \left[\begin{array}{rrrrr} -2 & 1 & 0 & 0 & 0 \\ \hline 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \\ \hline 0 & 0 & 0 & 1 & -2 \end{array}\right]. \tag{63}$$

In applying this operator to the data, the first and the last elements of $N_T y$, which are denoted by $n_0' y$ and $n_{T-1}' y$, respectively, are not true differences. Therefore, they are discarded to leave $-Q'y = [q_1, \ldots, q_{T-2}]'$. To compensate for this loss, appropriate values are attributed to q_0 and q_{T-1}, which are formed from combinations of the adjacent values, to create a vector of order T denoted by $q = [q_0, q_1, \ldots, q_{T-2}, q_{T-1}]'$.

The highpass filtering of the data comprises the following steps. First, the vector q is translated to the frequency domain to give $\gamma = Uq$. Then, the frequency-response matrix J_η is applied to the resulting Fourier ordinates. Next, in order to compensate for the effects of differencing, the vector of Fourier ordinates is premultiplied by a diagonal matrix $V = \mathrm{diag}\{v_0, v_1, \ldots, v_{T-1}\}$, wherein $v_j = 1/f(\omega_j); j = 0, \ldots, T-1$, with $\omega_j = 2\pi j/T$. Finally, the result is translated back to the time domain to create the vector h.

The vector of the complementary component is $x = y - h$. Thus there are

$$h = \bar{U}H_\eta U q \qquad \text{and} \qquad x = y - \bar{U}H_\eta U q, \tag{64}$$

where $H\eta = VJ_\eta$. It should be noted that the technique of re-inflating the data within the frequency domain cannot be applied in the case of a lowpass component for the reason that $f(0) = 0$ and, therefore, the function $v(\omega) = 1/f(\omega)$ is unbounded at the zero frequency $\omega = 0$. However, as the above equations indicate, this is no impediment to the estimation of the corresponding component x.

An example of the application of these procedures is provided by Figure 13, which concerns the familiar logarithmic consumption data, through which a smooth trend-cycle function has been interpolated. This is indistinguishable from the function that is obtained by adding the smooth business-cycle of Figure 12 to the linear trend that was subtracted from the data in the process of detrending it. The program also allows the trend-cycle function to be constructed in this manner.

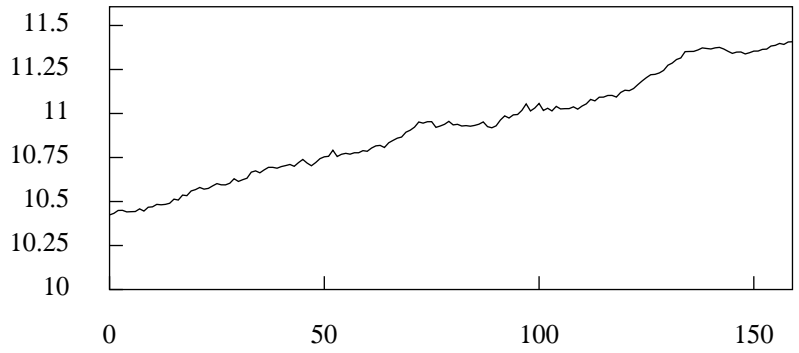


Fig. 14. The plot of a seasonally adjusted version of the consumption data of Figures 2 and 13, obtained via the time domain filter.

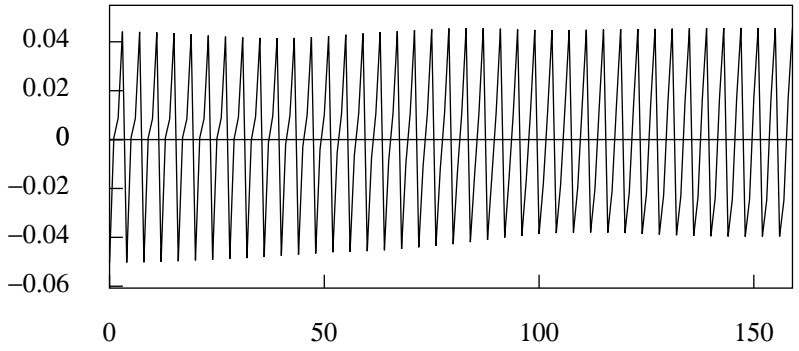


Fig. 15. The seasonal component extracted from the U.K. consumption data by a time-domain filter.

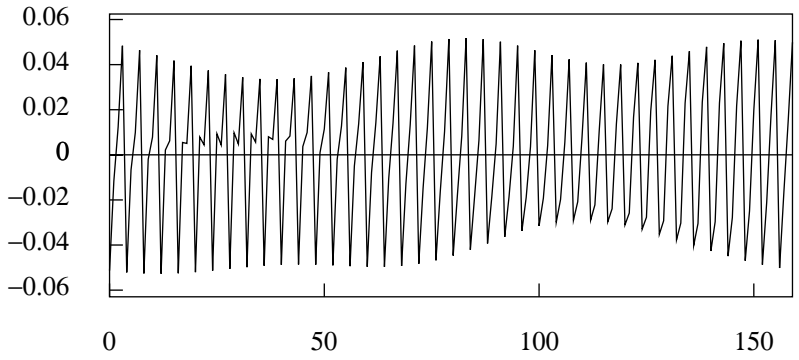


Fig. 16. The seasonal component extracted from the U.K. consumption data by a frequency-domain filter.

5.4 Seasonal Adjustment in the Frequency Domain

The method of frequency-domain filtering is particularly effective in connection with the seasonal adjustment of monthly or quarterly data. It enables one to remove elements not only at the seasonal frequencies but also at adjacent frequencies by allowing one to define a neighbourhood for each of the stop bands surrounding the fundamental seasonal frequency and its harmonics.

If only the fundamental seasonal element and its harmonics are entailed in its synthesis, then the estimated seasonal component will be invariant from year to year. If elements at the adjacent frequencies are also present in the synthesis, then it will evolve gradually over the length of the sample period.

The effects of the seasonal-adjustment filters of the program are illustrated in Figures 14–16. Figure 14 shows the seasonally adjusted version of the logarithmic consumption data that has been obtained via the Wiener–Kolmogorov filter of section 4. Figure 15 shows the seasonal component that has been extracted in the process.

The regularity of this component is, to some extent, the product of the filter. Figure 16 shows a less regular seasonal component that has been extracted by the frequency-domain filter described in the present section. This component has been synthesised from elements at the Fourier frequencies and from those adjacent to them that have some prominence if the periodogram of Figure 10.

6 The Program and its Code

The code of the program that has been described in this paper is freely available at the web address that has been given. This code is in Pascal. A parallel code

in C has been generated with the help of a Pascal-to-C translator, which has been written by the author. The aim has been to make the program platform-independent and to enable parts of it to be realised in other environments.

This objective has dictated some of the features of the user interface of the program, which, in its present form, eschews such devices as pull-down menus and dialogue boxes etc. Subsequent versions of the program will make limited use of such enhancements.

However, the nostrum that a modern computer program should have a modeless interface will be resisted. Whereas such an interface is necessary for programs such as word processors, where all of the functions should be accessible at all times, it is less appropriate to statistical programs where, in most circumstances, the user will face a restricted set of options. Indeed, the present program is designed to restrict the options, at each stage of the operations, to those that are relevant.

A consequence of this design is that there is no need of a manual of instructions to accompany the program. Instead, the three log files that record the steps taken in filtering some typical data sequences should provide enough help to get the user underway. What is more important is that the user should understand the nature of the statistical procedures that have been implemented; and this has been the purpose of the present paper.

Bibliography

[1] Baxter, M., and R.G. King, (1999). Measuring Business Cycles: Approximate Band-Pass Filters for Economic Time Series. *Review of Economics and Statistics,* 81, 575–593.

[2] Bloomfield, P., (1976). *Fourier Analysis of Time Series: An Introduction.* John Wiley and Sons, New York.

[3] Butterworth, S., (1930). On the Theory of Filter Amplifiers. *The Wireless Engineer* (From 1923 to 1930, the journal was called *Experimental Wireless and the Radio Engineer*), 7, 536–541.

[4] Christiano, L.J. and T.J. Fitzgerald, (2003). The Band-pass Filter. *International Economic Review,* 44, 435–465.

[5] Henderson, R., (1916). Note on Graduation by Adjusted Average. *Transactions of the Actuarial Society of America,* 17, 43–48.

[6] Henderson, R., (1924). A New Method of Graduation. *Transactions of the Actuarial Society of America,* 25, 29–40.

[7] Hodrick, R.J., and E.C. Prescott, (1980). *Postwar U.S. Business Cycles: An Empirical Investigation,* Working Paper, Carnegie–Mellon University, Pittsburgh, Pennsylvania.

[8] Hodrick, R.J., and E.C. Prescott, (1997). Postwar U.S. Business Cycles: An Empirical Investigation. *Journal of Money, Credit and Banking,* 29, 1–16.

[9] Ladiray, D., and B. Quenneville, (2001). *Seasonal Adjustment with the X-11 Method,* Springer Lecture Notes in Statistics 158, Springer Verlag, Berlin.

[10] Leser, C.E.V. (1961). A Simple Method of Trend Construction. *Journal of the Royal Statistical Society,* Series B, 23, 91–107.

[11] Musgrave, J. (1964). *A Set of End Weights to End all End Weights,* Working Paper, US Bureau of the Census, Washington

[12] Pollock, D.S.G., (2000). Trend Estimation and De-Trending via Rational Square Wave Filters. *Journal of Econometrics,* 99, 317–334.

[13] Pollock, D.S.G., (2007). Wiener–Kolmogorov Filtering, Frequency-Selective Filtering and Polynomial Regression. *Econometric Theory,* 23, 71–83.

[14] Pollock, D.S.G., (2008). Investigating Economic Trends and Cycles, in Palgrave Handbook of Econometrics: Vol. 2 Applied Econometrics, T.C. Mills and K. Patterson (editors). Palgrave Macmillan Ltd, Houndmills, Basingstoke.

[15] Quenneville, B., D. Ladiray and B. Lefranc, (2003). A Note on Musgrave Asymmetrical Trend-cycle Filters. *International Journal of Forecasting,* 19, 727–734.

[16] Whittle, P., (1983). *Prediction and Regulation by Linear Least-Square Methods,* Second Edition, Basil Blackwell, Oxford.

Who Uses gretl? An Analysis of the SourceForge Download Data

Riccardo (Jack) Lucchetti

Dipartimento di Economia - Università Politecnica delle Marche - Ancona, Italy
r.lucchetti@univpm.it

Abstract. This paper analyses the SourceForge download data to infer some characteristics of the population of gretl users. The rising number of downloads indicates Gretl's strong popularity as a teaching tool; however, despite the vast improvements in its features and performance, gretl's perceived status as a computational platform for research does not seem to be firmly established as yet, although this may change in the medium-long run.

1 Introduction

In the past few years, gretl has undoubtably come a long way in terms of features: thanks to constant feedback by a loyal user base and, most importantly, to Allin Cottrell's incredible committment and outstanding productivity, what used to be considered little more than a toy package now comprises a range of features which equal, and in some cases surpass, those found in commercial statistical programs. For example, the scope and efficiency of the routines for estimating GARCH-like models written for gretl by Balietti [2] are unrivalled by any other free software package.

The question I ask myself in this paper is: how has the gretl userbase evolved in response to the new features and the overall increase in usability of the package? In order to provide an answer, I will analyse the download data from SourceForge.

2 The SourceForge data

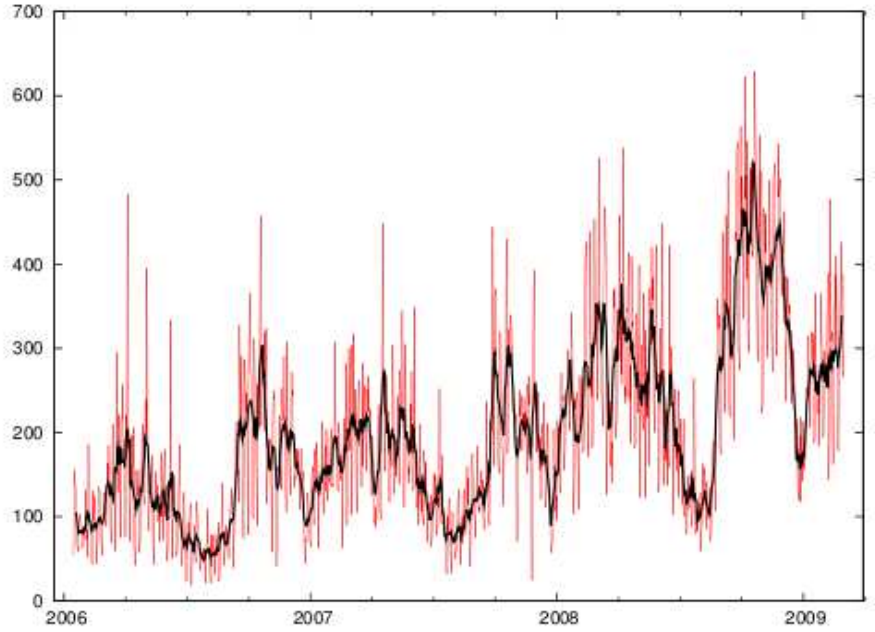
SourceForge is arguably the largest and most important hosting site for Free Software projects. It currently hosts tens of thousands of projects, among which hugely successful ones such as eMule and 7-zip.

Gretl ranks, at present, about 930th for number of total downloads, which amounts to about 300,000. However, it must be stressed that the number of downloads may not match the number of users for several reasons:

- The total number of downloads refers to all versions; a user who installed 10 versions of gretl on a computer counts as 10 downloads;

– some people may download the same version more than once, to install on different machines, possibly on different architectures;
– some people may download the "installer" once and give it to other people or make it available over a LAN;
– some gretl users may bypass SourceForge completely, especially those linux users who prefer to use the pre-packaged version of gretl offered by their linux distribution of choice (eg Debian and Ubuntu).

Fig. 1. Gretl daily downloads



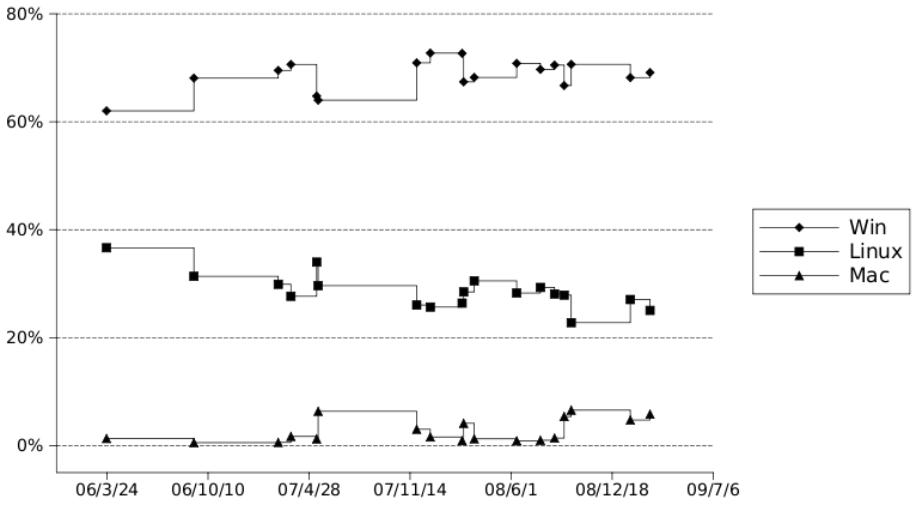
With this proviso, the number of daily downloads from SourceForge in the period 2006/1/15–2009/2/28 is shown in figure 1, together with a 7-day centered moving average. A few features are apparent:

– strong upward trend
– strong weekday effect
– strong seasonality

The upward trend is partly a consequence of the increase in popularity of Free Software at large, partly due to gretl's expanding user base (as confirmed by other anectodal evidence, such as the number of mailing list subscriptions

and so on); the weekday effect is unsurprising and of little interest in itself. The seasonal effect is, most likely, linked to the customary organisation of university courses. Roughly, the pattern is that downloads are low during the summer and the Christmas and Easter periods, while they spike up in September-October and in late February-March, which coincide with the beginning of terms in most universities, at least in the Northern hemisphere.

Fig. 2. Gretl downloads by architecture



It appears that gretl has reached, in three years, a much broader audience. This is, to some extent, confirmed by disaggregating the downloads of the various releases by platform.[1] Figure 2 shows the shares of downloads by platforms for each release from the beginning of 2006 to February 2009. The most striking feature of figure 2 is the decline of linux: I take this as evidence that gretl is now less perceived as a "geek" application than it used to be in 2006. Also notable is the increase of downloads for the Macintosh platform, which parallels the surge of its popularity among the general public[2].

Table 1. Weekly model

	Coefficient	Std. Error	t-ratio	p-value
const	2.82207	0.33705	8.3729	0.0000
y_1	0.38438	0.07337	5.2386	0.0000
tim	0.22222	0.03123	7.1154	0.0000
rel_1	0.11490	0.04764	2.4117	0.0171
a1	0.01668	0.02061	0.8093	0.4196
a2	0.13703	0.02479	5.5284	0.0000
s1	−0.01376	0.02257	−0.6094	0.5432
s2	−0.24451	0.03247	−7.5298	0.0000
q1	0.09449	0.02238	4.2213	0.0000
q2	0.03704	0.01959	1.8908	0.0606

Mean dependent var	5.168366	S.D. dependent var	0.499487
Sum squared resid	4.522887	S.E. of regression	0.172499
R^2	0.887399	Adjusted R^2	0.880732
$F(9, 152)$	133.1000	P-value(F)	1.99e–67
Log-likelihood	59.98606	Akaike criterion	−99.97212
Schwarz criterion	−69.09616	Hannan–Quinn	−87.43601
$\hat{\rho}$	−0.006229	Durbin's h	−0.216540

LM test for autocorrelation up to order 7 –
  Test statistic: LMF = 1.16168
  with p-value = $P(F(7, 145) > 1.16168)$ = 0.328463

Test for ARCH of order 1 –
  Test statistic: LM = 2.85511
  with p-value = $P(\chi^2(1) > 2.85511)$ = 0.091084

Koenker test for heteroskedasticity –
  Test statistic: LM = 7.84029
  with p-value = $P(\chi^2(9) > 7.84029)$ = 0.550318

Test for normality of residual –
  Test statistic: $\chi^2(2)$ = 1.94429
  with p-value = 0.378271

QLR test for structural break –
  Test statistic: $F_{\max}(10, 142)$ = 2.12638 (06/11/06)
  (10 percent critical value = 2.48)

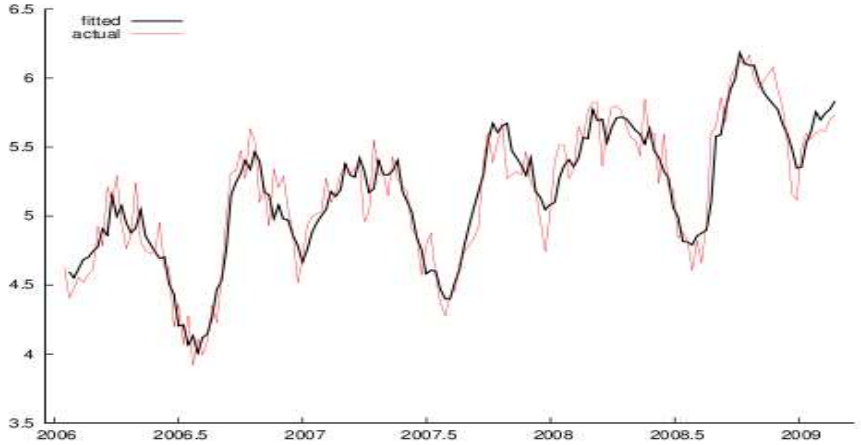
3 A model for weekly downloads

In this section, I will analyse downloads (in logs), after aggregating the data on a weekly basis, to get rid of the weekday effect. The explanatory variables are a time trend and a combination of sine-cosine terms with annual, semi-annual and quarterly period.[3] Additional regressors to account for short-term fluctuations are a dummy variable for the emergence of a new release on the previous week and one lag of the dependent variable. The model can therefore be represented as

$$(1 - \phi L)y_t = \beta_0 + \beta_1 t + \beta_2 r_{t-1} + \gamma' s_t + \varepsilon_t \tag{1}$$

where s_t is a vector of six trigonometric terms.

Fig. 3. Actual and fitted weekly downloads (logs)



OLS estimates of equation (1) are presented in table 1. As can be seen, the fit is excellent. The model predicts a rate of growth of about 43% per year[4] and the seasonal effect, as captured by the trigonometric terms, is highly significant.

[1] Downloads of the source package were assimilated to other linux packages.

[2] According to the Marketshare website (`http://marketshare.hitslink.com/`), the share of Mac users on the Net has risen from 6.09% in March 2007 to 9.61% in February 2009.

[3] Higher frequencies were also tried, but turned out to provide no significant contribution.

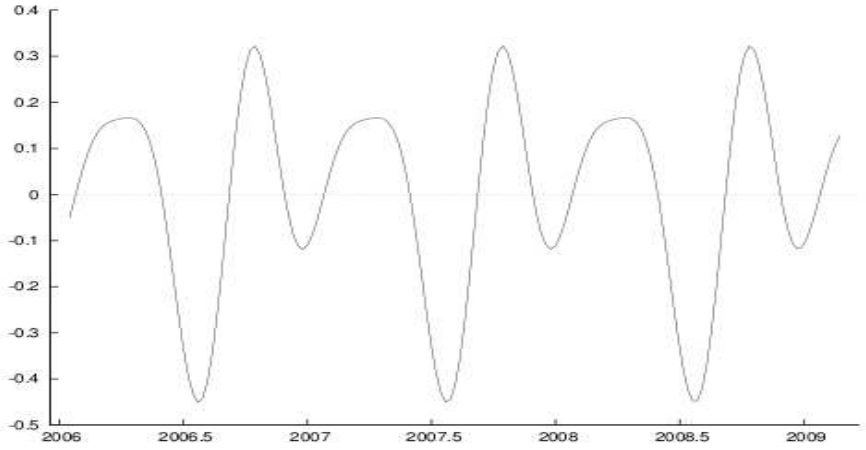
[4] This is computed as

$$\exp\left(\frac{\hat{\beta}_1}{1-\hat{\phi}}\right) - 1.$$

Moreover, the pure seasonal component (plotted in figure 4) shows clearly the summer and Christmas slowdowns I mentioned earlier. When a new version of gretl is released, downloads rise in the following week by about 12%.

Finally, the customary diagnostic tests show no sign of mis-specification, so we can conclude that the above model is adequate in summarising the data and that the two main stylised facts (the increase in downloads and its seasonality) are robust. The output of the Quandt LR test is especially important, since it indicates that the time pattern of gretl downloads has remained reasonably stable through our three-year sample.

Fig. 4. Seasonal component



4 Research or teaching?

The above model can be used for an indirect analysis of the composition of gretl users (or, more correctly, downloaders) between "researchers" and "teaching people" (which include students and teachers).

Clearly, this distinction is, to some extent, spurious: I, for one, belong to both categories. However, while gretl's aptness as a teaching tool is widely acknowledged (see for example Smith and Mixon [6] or Adkins [1]), there seems to be little recognition of gretl as a tool for applied research;[5] notable exceptions

[5] Under-reporting may be an issue here; research papers seldom cite the software used for their computations: for example, a recent paper of mine (Lucchetti and Palomba [4]) makes no mention of gretl whatsoever, despite the fact that gretl was almost exclusively used.

are Yalta and Yalta [7] (now slightly dated) and Rosenblad [5], who discusses gretl as a teaching tool but also highlights its potential for research use.

In my opinion, the composition of gretl's user base is of crucial importance for shaping gretl's future: people from the "teaching" community provide invaluable feedback and motivation to keep improving gretl's already excellent user interface. On the other hand, gretl's computing accuracy, scope and efficiency are put to the test (and pushed forward accordingly) primarily by those who use it for "real" applied statistical tasks. Moreover, if a true community of coders is to emerge, it is more likely that new coders come from the ranks of young, computer-literate researchers.[6] Therefore, it is important to ascertain the composition of gretl's user base if one wants to forecast, and possibly steer, the evolution of gretl as a computing platform.

Of course, a few assumptions are necessary here: I assume (a) that research activity is not seasonal and (b) that the share of "researchers" is a linear function of time. Needless to say, both hypotheses are a bit strong; we all have holidays in the summer and even atheists take a Christmas break. On the other hand, most of us do take advantage of periods when classes stop, to work on our research papers.

However, assumption (b) should be also only viewed as an approximation, meant to capture the overall trend, rather than a hard-edged fact and is in itself rather innocuous. As an indirect confirmation of assumption (a), I ran a simple seasonality extraction procedure on monthly download data for four popular scientific applications[7]: apart from gretl, I used gnuplot (the data visualistion tool that gretl also uses), maxima (a computer algebra system) and the GNU Octave software repository (a collection of additional modules for GNU octave, a free Matlab replacement). All these are mature projects, which are certainly used in teaching but also have a well-established reputation as research tools.

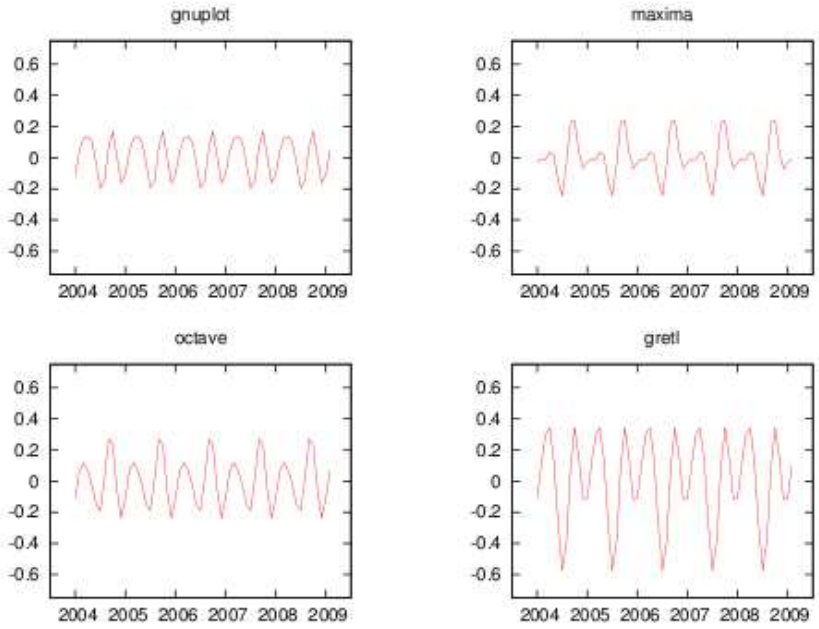
For each of these packages, the seasonality extraction routine is based on an OLS regression of the log of monthly downloads on a constant, one lag, a linear trend and the six trigonometric variables used above. The seasonality component is then reconstructed as the sum of each trigonometric variable weighted by its own estimated coefficient. Results are plotted in figure 5: it should be obvious that gretl's seasonal component is much larger than the other packages'.

In the light of assumptions (a) and (b), equation (1) can be generalised via a logistic smooth-transition regression model, similar in spirit to a STAR model

[6] It is true, however, that having been exposed to gretl as a student may encourage a researcher to study the source code and contribute original ideas and code.

[7] The source in all cases is, again, SourceForge. Unfortunately, I was unable to find download data for the R statistical project, which is not hosted by SourceForge and would have been extremely interesting to analyse.

Fig. 5. Seasonal component on monthly data for several free software projects



(see Granger and Teräsvirta [3]):

$$(1 - \phi L)y_t = \beta_0 + \beta_1 t + \beta_2 r_{t-1} + \frac{2}{1 + \exp(-\alpha t)}\gamma' s_t + \varepsilon_t \qquad (2)$$

Here, the α parameter measures the time variation of the importance of the seasonal component: if assumption (a) above is valid, then the basic idea is that α is a rough measure of how the weight of "teaching people" on the whole gretl ecosystem increases through time. Put another way, if α is greater than (less than) 0, then the share of people who download gretl for research decreases (increases). If α equals 0, the model reduces to (1).

Equation (2) was estimated by nonlinear least squares: the estimation results are shown in table 2. As is apparent, the coefficients are roughly the same as those in table 1. The estimate of α is negative, which suggests a reduction in time of the seasonal component, but is far from being significant. Hence, there is no compelling evidence of a reduction of the importance of the seasonality component in gretl downloads. If assumption (a) is valid, this means that in the period 2006-2008 the fraction of gretl downloads for teaching purposes has remained more or less stable.

Two considerations must be made at this point: first, the adoption of a statistical package as the tool of choice by applied economists and econometricians is

Table 2. Weekly nonlinear model

	Estimate	Std. Error	t-ratio	p-value
const	2.82425	0.33827	8.349	0.0000
y_1	0.38382	0.07364	5.212	0.0000
tim	0.11411	0.04787	2.384	0.0184
rel_1	-0.22265	0.03134	-7.104	0.0000
a1	0.01663	0.02139	0.778	0.4380
a2	0.14188	0.02995	4.738	0.0000
s1	-0.01456	0.02357	-0.618	0.5377
s2	-0.25353	0.04499	-5.635	0.0000
q1	0.09868	0.02715	3.634	0.0004
q2	0.03803	0.02071	1.836	0.0683
α	-0.04465	0.14856	-0.301	0.7642

Mean dependent var	5.168366	S.D. dependent var	0.499487
Sum squared resid	4.520373	S.E. of regression	0.173021
R^2	0.887462	Adjusted R^2	0.880009
Log-likelihood	60.03124	Akaike criterion	−98.06221
Schwarz criterion	−64.09892	Hannan–Quinn	−84.27276
$\hat{\rho}$	−0.005810	Durbin–Watson	2.001861

a long process: path-dependence and acquired habits may cause people to stick to obsolete tools for years, so, even if assumptions (a) and (b) are valid, it may just be the case that the sample I am using here is simply too short to capture this aspect adequately.

Moreover, the emergence of a community of gretl code contributors, well-versed in econometrics and programming at the same time, is unlikely to depend on the relative share of researchers on gretl's total user base, but rather on its absolute value. What counts is the number of "hackers" we have, not the percentage of users who are. In this sense, the stability of the share of "research people" on an increasing number of users allows us to be mildly optimistic.

5 Conclusions

Gretl has been so far a spectacular success story in terms of expansion of its user base. Obviously, the characteristic of free software that most people perceive as paramount (being "free as in beer") played its role, but this factor does not explain the whole story by itself: the Internet is full of gratis software which few people, if any, use. A large part of the merit goes to its intuitive and friendly interface, which makes it an ideal tool for teaching econometrics.

On the other hand, gretl's capabilities as a computing platform, which have also evolved dramatically, seem to have been overlooked by most practitioners, although lack of evidence may simply be a consequence of the limited time span of the data used here. Only time will tell gretl's future reputation as a solid and reliable computing nad there are reasons for moderate optimism. In any case, it is vitally important for the gretl community to work to advertise gretl's present capabilities to the widest possible audience and to work as hard as possible to extend and perfect them.

Bibliography

[1] ADKINS, L. (2009): *Using gretl for Principles of Econometrics, 3rd edition*. online.

[2] BALIETTI, S. (2008): "gig: a GARCH-like Implementation in Gretl," Master's thesis, Università Politecnica delle Marche.

[3] GRANGER, C. W., AND T. TERÄSVIRTA (1993): *Modelling Nonlinear Economic Relationships*. Oxford University Press.

[4] LUCCHETTI, R., AND G. PALOMBA (2009): "Nonlinear adjustment in US bond yields: An empirical model with conditional heteroskedasticity," *Economic Modelling*, forthcoming.

[5] ROSENBLAD, A. (2008): "gret1 1.7.3," *Journal of Statistical Software, Software Reviews*, 25(1), 1–14.

[6] SMITH, R. J., AND J. W. MIXON (2006): "Teaching undergraduate econometrics with GRETL," *Journal of Applied Econometrics*, 21(7), 1103–1107.

[7] YALTA, A. T., AND A. Y. YALTA (2007): "GRETL 1.6.0 and its numerical accuracy," *Journal of Applied Econometrics*, 22(4), 849–854.

Econometric Theory

An Instrumental Variables Probit Estimator using gretl

Lee C. Adkins

Professor of Economics, Oklahoma State University, Stillwater, OK 74078
lee.adkins@okstate.edu

Abstract. The most widely used commercial software to estimate endogenous probit models offers two choices: a computationally simple generalized least squares estimator and a maximum likelihood estimator. Adkins [1, 2] compares these estimators to several others in a Monte Carlo study and finds that the GLS estimator performs reasonably well in some circumstances. In this paper the small sample properties of the various estimators are reviewed and a simple routine using the gretl software is given that yields identical results to those produced by Stata 10.1. The paper includes an example estimated using data on bank holding companies.

1 Introduction

Yatchew and Griliches [19] analyze the effects of various kinds of misspecification on the probit model. Among the problems explored was that of errors-in-variables. In linear regression, a regressor measured with error causes least squares to be inconsistent and Yatchew and Griliches find similar results for probit. Rivers and Vuong [14] and Smith and Blundell [16] suggest two-stage estimators for probit and tobit, respectively. The strategy is to model a continuous endogenous regressor as a linear function of the exogenous regressors and some instruments. Predicted values from this regression are then used in the second stage probit or tobit. These two-step methods are not efficient, but are consistent. Consistent estimation of the standard errors is not specifically considered and these estimators are used mainly to test for endogeneity of the regressors–not to establish their statistical significance.

Newey [12] looked at the more generic problem of endogeneity in limited dependent variable models (which include probit and tobit). He proposed what is sometimes called Amemiya's Generalized Least Squares (AGLS) estimator as a way to efficiently estimate the parameters of probit or tobit when they include a continuous endogenous regressor. This has become one of the standard ways to estimate these models and is an option (twostep) in Stata 10.0 when the MLE is difficult to obtain. The main benefit of using this estimator is that it yields a consistent estimator of the variance-covariance matrix that and can easily be used for subsequent hypothesis tests about the parameters.

Adkins [1] compares the AGLS estimator to several alternatives, which include a maximum likelihood estimator. The AGLS estimator is simple to compute and yields significance tests that are close in size to the nominal level when samples are not very large (e.g., n=200). The other plug-in estimators are consistent for the parameters but not the standard errors, making it unlikely that they will perform satisfactorily in hypothesis testing. The latter problem is taken up by Adkins [3] who uses a Murphy and Topel [11] correction to obtain consistent standard errors with some success.

Others have explored limited dependent variable models that have discrete endogenous regressors. Nicoletti and Peracchi [13] look at binary response models with sample selection, Kan and Kao [10] consider a simulation approach to modeling discrete endogenous regressors, and Arendt and Holm [5] extends Nicoletti and Peracchi [13] to include multiple endogenous discrete variables.

Iwata [9] uses a very simple approach to dealing with errors-in-variables for probit and tobit. He shows that simple recentering and rescaling of the observed dependent variable may restore consistency of the standard IV estimator if the true dependent variable and the IV's are jointly normally distributed. His Monte Carlo simulation shows evidence that the joint normality may not be necessary to obtain improved results. However, the results for tobit were quite a bit better than those for probit. He compares this estimator to a linear instrumental variable estimator that uses a consistent estimator of standard errors. This estimator is considered by Adkins [1] in his comparison.

Blundell and Powell [6] develop and implement semiparametric methods for estimating binary dependent variable models that contain continuous endogenous regressors. Their paper "extends existing results on semiparametric estimation in single-index binary response models to the case of endogenous regressors. It develops an approach to account for endogeneity in triangular and fully simultaneous binary response models." Blundell and Powell [6], p. 655

In the following sections a linear model with continuous endogenous regressors and its estimators are considered. With respect to models having a dichotomous dependent variable, a relatively simple generalized least squares estimator discussed in Newey [12] is presented and an algorithm for its computation in gretl is given. To give the reader an idea of how this estimator compares to alternatives, including a maximum likelihood estimator (mle), some results from a simulation study conducted by Adkins [1, 2] are summarized. The results from the gretl routine and from Stata 10 are compared using an example from the banking literature.

2 Linear Model and Estimators

Following the notation in Newey [12], consider a linear statistical model in which the continuous dependent variable will be called y_t^* but it is not directly observed. Instead, we observe y_t in only one of two possible states. So,

$$y_t^* = Y_t\beta + X_{1t}\gamma + u_t = Z_t\delta + u_t, \quad t = 1, \ldots, N \tag{1}$$

where $Z_t = [Y_t, X_{1t}], \delta' = [\beta', \gamma']$, Y_t is the t^{th} observation on an endogenous explanatory variable, X_{1t} is a 1xs vector of exogenous explanatory variables, and δ is the $qx1$ vector of regression parameters.

The endogenous variable is related to a $1Xk$ vector of instrumental variables X_t by the equation

$$Y_t = X_{1t}\Pi_1 + X_{2t}\Pi_2 + V_t = X_t\Pi + V_t \tag{2}$$

where V_t is a disturbance. The $k - s$ variables in X_{2t} are additional exogenous explanatory variables. Equation (2) is the reduced form equation for the endogenous explanatory variable. Without loss of generality only one endogenous explanatory variable is considered below. See Newey [12] for notation extending this to additional endogenous variables.

When the continuous variable y_t^* is observed, then one could use either least squares or instrumental variable estimator to estimate δ. Collecting the n observations into matrices y^*, X, and Z of which the t^{th} row is y_t^*, X_t, and Z_t, respectively we have the least squares estimator of δ, $\hat{\delta}_{ols} = (Z^TZ)^{-1}Z^Ty^*$, which is biased and inconsistent.

The instrumental variable estimator uses the orthogonal projection of Z onto the column space of X, i.e., P_XZ where $P_X = X(X^TX)^{-1}X^T$. The IV estimator is

$$\delta_{liv} = (Z^TP_XZ)^{-1}Z^TP_Xy^*. \tag{3}$$

The (linear) instrumental variable estimator is biased in finite samples, but consistent. The heteroskedasticity robust estimator of covariance Davidson and MacKinnon [7], p. 335 is

$$\hat{\Sigma}_{HCCME} = (Z^TP_XZ)^{-1}Z^TP_X\hat{\Phi}P_XZ(Z^TP_XZ)^{-1} \tag{4}$$

where $\hat{\Phi}$ is an nxn diagonal matrix with the t^{th} diagonal element equal to $\hat{u}_t^2$, the squared IV residual.

3 Binary Choice Model and Estimators

In some cases, y_t^* is not directly observed. Instead, we observe

$$y_t = \begin{cases} 1 & y_t^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Assuming the errors of the model (1) are normally distributed leads to the probit model.

3.1 Linear, MLE, and Plug-in

There are several estimators of this model, some consistent for δ and others not. The first is least squares. The least squares estimator $\hat{\delta}_{ols} = (Z^T Z)^{-1} Z^T y^*$ is consistent if Z is exogenous. If any of the elements of Z are endogenous then it is not. Still, it is easy to compute and the degree of inconsistency may be small in certain circumstances.

The linear instrumental variable estimator (3) is also inconsistent and heteroscedastic. Iwata [9] suggests a means of rescaling and recentering (RR) the data that can bring about consistency in this case. However, in his Monte Carlo the RR versions of OLS and IV estimation don't perform particularly well for probit (although much better for tobit).

The usual probit mle can be used to estimate the parameters. However, when any of the regressors are endogenous, then this estimator is also inconsistent (Yatchew and Griliches [19]). To develop the notation, let the probability that y_t is equal one be denoted

$$pr(y_t = 1) = \Phi(y_t, Y_t\beta + X_{1t}\gamma) = \Phi(y_t, Z_t\delta) \tag{6}$$

where Φ is the normal cumulative density, y_t is the observed binary dependent variable, and $Y_t\beta + X_{1t}\gamma$ is the (unnormalized) index function. As usual, the model is normalized assuming $\sigma^2 = 1$. Basically, this equation implies that Y_t, and X_{1t} be included as regressors in the probit model and the log likelihood function is maximized with respect to $\delta^T = [\beta^T, \gamma^T]$. Since the endogeneity of Y_t is ignored, the mle is inconsistent.

Another estimator uses predicted values of Y_t from a first stage least squares estimation of equation (2). Denote the first stage as $\hat{Y}_t = X_{1t}\hat{\Pi}_1 + X_{2t}\hat{\Pi}_2 = X_t\hat{\Pi}$ where $X_t = [X_{1t} \vdots X_{2t}]$ and $\hat{\Pi}^T = [\hat{\Pi}_1^T \vdots \hat{\Pi}_2^T]$. Then the conditional probability

$$pr(y_t = 1) = \Phi(y_t, \hat{Z}_t\delta) \tag{7}$$

with $\hat{Z}_t = [\hat{Y}_t \vdots X_{1t}]$. The parameters are found by maximizing the conditional likelihood. This is referred to here as IV probit (IVP). Although IVP is consistent for δ the standard errors estimated as the outer product of the gradient are not. This can be easily remedied using a Murphy and Topel [11] type correction.

Another estimator adds the least squares residuals from equation (2), $\hat{V}_t = Y_t - X_t\hat{\Pi}$ to (7). This brings

$$pr(y_t = 1) = \Phi(y_t, \hat{Y}_t\beta + X_{1t}\gamma + \hat{V}_t\lambda) = \Phi(y_t, \hat{Z}_t\delta + \hat{V}_t\lambda) \tag{8}$$

which is estimated by maximum likelihood, again conditional on $\hat{\Pi}$. This is similar to an estimator used by Rivers and Vuong [14] which takes the form

$$pr(y_t = 1) = \Phi(y_t, Z_t\delta + \hat{V}_t\rho) \tag{9}$$

The parameter ρ is related to λ in (8) by $\lambda = \rho + \beta$. This follows because $Z_t\delta = \hat{Z}_t\delta + \hat{V}_t\beta$. This estimator is useful for testing endogeneity, but seldom used to estimate δ.

3.2 AGLS

An efficient alternative to (8), proposed by Newey [12], and credited to Amemiya, is a generalized least squares estimator (AGLS). The AGLS estimator of the endogenous probit model is fairly easy to compute, though there are several steps–more than the two suggested by the name of its option in Stata. The basic algorithm proceeds as follows:

1. Estimate the reduced form (2), saving the estimated residuals, $\hat{V}_t$ and predicted values $\hat{Y}_t$.
2. Estimate the parameters of a reduced form equation for the probit model using the mle. In this case,

   $$pr(y_t = 1) = \Phi(y_t, X_t\alpha + \hat{V}_t\lambda) \tag{10}$$

   Note that all exogenous variables, X_{1t} and instruments X_{2t} are used in the probit reduced form and the parameters on these variables is labeled α. Let the mle be denoted $\hat{\alpha}$. Also, save the portion of the estimated covariance matrix that corresponds to $\hat{\alpha}$, calling it $\hat{J}^{-1}_{\alpha\alpha}$.
3. Another probit model is estimated by maximum likelihood. In this case it is the 2SIV estimator of equation (8). Save $\hat{\rho} = \hat{\lambda} - \hat{\beta}$ which is the coefficient of $\hat{V}_t$ minus that of $\hat{Y}_t$.
4. Multiply $\hat{\rho}Y_t$ and regress this on X_t using least squares. Save the estimated covariance matrix from this, calling it $\hat{\Sigma}$.

5. Combine the last two steps into a matrix, $\Omega = \hat{J}_{\alpha\alpha}^{-1} + \hat{\Sigma}$.
6. Then, the AGLS estimator is

$$\delta_A = [D(\hat{\Pi})^T \Omega^{-1} D(\hat{\Pi})]^{-1} D(\hat{\Pi})^T \Omega^{-1} \hat{\alpha} \tag{11}$$

The estimated variance covariance is $[D(\hat{\Pi})^T \Omega^{-1} D(\hat{\Pi})]^{-1}$ and $D(\hat{\Pi}) = [\hat{\Pi} \vdots I_1]$ where I_1 is a kxs selection matrix such that $X_{1t} = X_t I_1$.

The AGLS estimator is one of the options available in Stata 10 (the other is an mle). Adkins [2, 1] conducts a Monte Carlo simulation to compare the bias of each of these estimators as well as the size of nominal 10% significance test of model parameter. He finds that in some circumstances the AGLS estimator performs reasonably well and can be used successfully to test for significance, especially if the sample is small and the instruments not very strong. The main findings of Adkins [1] are reproduced in the next section.

4 Summary of Simulation Results from Adkins (2008)

The main results of Adkins [1] can be summarized as follows:

1. When there is no endogeneity, OLS and Probit work well (as expected). Bias is very small and tests have the desired size.
2. It is clear that OLS and Probit should be avoided when you have an endogenous regressor. Both estimators are significantly biased and significance tests do not have the desired size.
3. Weak instruments increases the bias of AGLS. The bias worsens as the correlation between the endogenous regressor and the equation's error increases.
4. The actual size of a parameter significance test based on the instrumental variable probit is reasonably close to the nominal level in nearly every instance. This is surprising for at least two reasons. 1) The bias of IVP is substantial when instruments are weak. 2) The test statistic is based on an inconsistent estimator of the standard error. No attempt was made to estimate the covariance of this estimator consistently, as is done in Limdep 9 Greene [8]. This is explored further in Adkins [3] who uses a Murphy and Topel [11] correction to obtain consistent standard errors.
5. The size of the significance tests based on the AGLS estimator is also reasonable, but the actual size is larger than the nominal size–a situation that gets worse as severity of the endogeneity problem increases. When instruments are very weak, the actual test rejects a true null hypothesis twice as often as it should.

6. Linear instrumental variables estimators that use consistent estimators of standard errors can be used for this purpose (significance testing) though their performance is not quite up to that of the AGLS estimator. The Linear IV estimator performs better when the model is just identified.
7. There is an improvement in bias and the size of the significance test when samples are larger (n=1000). Mainly, smaller samples (n=200) require stronger instruments in order for bias to be small and tests to work properly (other than IVP, which as mentioned above, works fairly well all the time).
8. There is little to be gained by pretesting for endogeneity. When instruments are extremely weak it is outperformed by the other estimators considered, except when the no endogeneity hypothesis is true (and probit should be used). Bias is reduced by small amounts, but it is uncertain what one would use as an estimator of standard errors for a subsequent t-test.
9. When instruments are weak, t-tests based on ML are no better than ones based on AGLS (in fact, one could argue that they are worse). Significance testing based on the ML estimator is much more reliable in large samples.

The picture that emerges from this is that the AGLS estimator may be useful when the sample is relatively small and the instruments not very strong. It is also useful when the mle cannot be computed–a situation which limited the simulations conducted by Adkins [1, 2]. Given the usefulness of the AGLS estimator, a gretl script is provided to compute is and its standard errors. The script is provided below in section 6. In the next section, a brief example is given the results from Stata 10 and the gretl script are compared.

5 Example

In this section a brief example based on Adkins et al. [4] is presented and the results from Stata and gretl compared.

The main goal of Adkins et al. [4] was to determine whether managerial incentives affect the use of foreign exchange derivatives by bank holding companies (BHC). There was some speculation that several of the variables in the model were endogenous. The dependent variable of interest is an indicator variable that takes the value 1 if the BHC uses foreign exchange derivative. The independent variables are as follows:

Ownership by Insiders When managers have a higher ownership position in the bank, their incentives are more closely aligned with shareholders so they have an incentive to take risk to increase the value of the call option associated with equity ownership. This suggests that a higher ownership position by

insiders (officers and directors) results in less hedging. The natural logarithm of the percentage of the total shares outstanding that are owned by officers and directors is used as the independent variable.

Ownership by Institutional Blockholders Institutional blockholders have incentive to monitor the firm's management due to the large ownership stake they have in the firm (Shleifer and Vishny [15]). Whidbee and Wohar [18] argue that these investors will have imperfect information and will most likely be concerned about the bottom line performance of the firm. The natural logarithm of the percentage of the total shares outstanding that are owned by all institutional investors is included as an independent variable and predict that the sign will be positive, with respect to the likelihood of hedging.

CEO Compensation CEO compensation also provides its own incentives with respect to risk management. In particular, compensation with more option-like features induces management to take on more risk to increase the value of the option (Smith and Blundell [16]; Tufano [17]). Thus, higher options compensation for managers results in less hedging. Two measures of CEO compensation are used: 1) annual cash bonus and 2) value of option awards.

There is a possibility that CEO compensation is endogenous in that successful hedging activity could in turn lead to higher executive compensation. The instruments used for the compensation variables are based on the executive's human capital (age and experience), and the size and scope of the firm (number of employees, number of offices and subsidiaries). These are expected to be correlated with the CEOs compensation and be predetermined with respect to the BHCs foreign exchange hedging activities.

BHC Size The natural logarithm of total assets is used to control for the size of the BHC.

Capital The ratio of equity capital to total assets is included as a control variable. The variable for dividends paid measures the amount of earnings that are paid out to shareholders. The higher the variable, the lower the capital position of the BHC. The dividends paid variable is expected to have a sign opposite that of the leverage ratio.

Like the compensation variables, leverage should be endogenously determined. Firms that hedge can take on more debt and thus have higher leverage, other things equal.

Foreign Exchange Risk A bank's use of currency derivatives should be related to its exposure to foreign exchange rate fluctuations. The ratio of interest income from foreign sources to total interest income measures foreign exchange exposure. Greater exposure, as represented by a larger proportion of income being derived from foreign sources, should be positively related to both the likelihood and extent of currency derivative use.

Profitability The return on equity is included to represent the profitability of the BHCs. It is used as a control.

5.1 Results

In this section the results of estimation are reported. Table 1 contains some important results from the reduced form equations. Due to the endogeneity of leverage and the CEO compensation variables, instrumental variables estimation is used to estimate the probability equations. Table 2 reports the coefficient estimates for the instrumental variable estimation of the probability that a BHC will use foreign exchange derivatives for hedging. The first column of results correspond to the Stata two-step estimator and the second column, gretl.

In Table 1 summary results from the reduced form are presented. The columns contain p-values associated with the null hypothesis that the indicated instrument's coefficient is zero in each of the four reduced form equations. The instruments include number of employees, number of subsidiaries, number of offices, CEO's age–which proxies for his or her experience, the 12 month maturity mismatch, and the ratio of cash flows to total assets (CFA). The p-values associated with the other variables have been suppressed to conserve space.

Each of the instruments appears to be relevant in that each is significantly different from zero at the 10% (p-value < 0.1) in at least one equation; the number of employees, number of subsidiaries, and CEO age and CFA are significant in one equation; the number of offices, employees, subsidiaries are significant in two equations.

The overall strength of the instruments can be roughly gauged by looking at the overall fit of the equations. The R^2 in the leverage equation is the smallest (0.29), but is still high relative to the results of the Monte Carlo simulation. The instruments, other than the 12 month maturity mismatch, appear to be strong and we have no reason to expect poor performance from either the AGLS or the mle in terms of bias.

The simulations from Adkins [1] suggest discarding extra instruments, and this would be recommended here. Which to drop, other than the mismatch variable is unclear. CFA, Age, and subsidiaries are all strongly correlated with lever-

age; office and employees with options; and, employees, subsidiaries, and offices with bonuses. The fit in the leverage equation is weakest, yet the p-values for each individual variable is relatively high. For illustrative purposes, I'll plow forward with the current specification.

Table 1. Summary Results from Reduced-form Equations. The table contains p-values for the instruments and R^2 for each reduced form regression. The data are taken from the Federal Reserve System's Consolidated Financial Statements for Bank Holding Companies (FR Y-9C), the *SNL Executive Compensation Review*, and the *SNL Quarterly Bank Digest*, compiled by SNL Securities.

	Reduced Form Equation		
	Leverage	Options	Bonus
Instruments	Coefficient P-values		
Number of Employees	0.182	0.000	0.000
Number of Subsidiaries	0.000	0.164	0.008
Number of Offices	0.248	0.000	0.000
CEO Age	0.026	0.764	0.572
12 Month Maturity Mismatch	0.353	0.280	0.575
CFA	0.000	0.826	0.368
R-Square	0.296	0.698	0.606

Table 2: **IV Probit Estimates of the Probability of Foreign-Exchange Derivatives Use By Large U.S. Bank Holding Companies (1996-2000)**. This table contains estimates for the probability of foreign-exchange derivative use by U.S. bank holding companies over the period of 1996-2000. To control for endogeneity with respect to compensation and leverage, we use an instrumental variable probit estimation procedure. The dependent variable in the probit estimations (i.e., probability of use) is coded as 1 if the bank reports the use of foreign-exchange derivatives for purposes other than trading. The data are taken from the Federal Reserve System's Consolidated Financial Statements for Bank Holding Companies (FR Y-9C), the *SNL Executive Compensation Review*, and the *SNL Quarterly Bank Digest*, compiled by SNL Securities. Approximate p-values based on the asymptotic distribution of the estimators are reported in parentheses beneath the parameter estimates.

	Instrumental Variables Probit	
	Stata (twostep)	gretl
Leverage	21.775	21.775
	(13.386)	(13.386)
Option Awards	-8.79E-08	-8.79E-08
	(5.31E-08)	(5.31E-08)
Bonus	1.76E-06	1.76E-06

Continued from preceding page		
	Instrumental Variables Probit	
	Stata	gretl
	(8.88E-07)	(8.88E-07)
Total Assets	0.36453	0.36453
	(0.17011)	(0.17011)
Insider Ownership %	0.25882	0.25882
	(0.11623)	(0.11623)
Institutional Ownership %	0.36981	0.36981
	(0.13477)	(0.13477)
Return on Equity	-0.033852	-0.033852
	(0.028188)	(0.028188)
Market-to-Book ratio	-0.0018722	-0.0018722
	(0.0012422	(0.0012422)
Foreign to Total Interest Income Ratio	-3.5469	-3.546958
	(3.8414)	(3.8414)
Derivative Dealer Activity Dummy	-0.2799	-0.2799
	(0.24675)	(0.24675)
Dividends Paid	-8.43E-07	-8.43E-07
	(5.62E-07)	(5.62E-07)
D=1 if 1997	-0.024098	-0.024098
	(0.27259)	(0.27259)
D=1 if 1998	-0.24365	-0.24365
	(0.26195)	(0.26195)
D=1 if 1999	-0.24156	-0.24156
	(0.28171)	(0.28171)
D=1 if 2000	-0.128	-0.127999
	(0.27656)	(0.27656)
Constant	-9.673	-9.673
	(2.5351)	(2.5351)
Sample size	794	794

The model is overidentified, the sample is large (700+), and the instruments are very strong. Compared to maximum likelihood (ML) estimation, a few differences were found (see Adkins [2]). Leverage is significant in ML at the 10% level, but not with AGLS. Similarly, return-on-equity, market-to-book, and dividends paid are all significant in the ML regression but not AGLS. This divergence of results is a little troubling. In terms of the small sample properties documented by Adkins [1], ML p-values tend to be too small when instruments were mildly strong and correlation low. If the endogeneity problem is not severe, then the ML estimation and AGLS results tend to diverge. In this case, then AGLS estimator appears to be more reliable for testing significance. In the case of very strong instruments, the AGLS estimator tended to be insignificant too often. In the banking example, the empirical model falls between these two extremes and a strong recommendation can not be made for one over the other.

However, for the purposes of this paper, the news is excellent: the Stata results (column 1) and those from the simple gretl script (column 2) are basically identical. In situations where the AGLS is called for, one can confidently use the gretl script provided below to estimate the parameters of probit model that contains continuous endogenous regressors.

6 gretl Script

The following script was used with gretl 1.7.8 to produce the results in column 2 of Table 2.

```
# Variable definitions
# y2 = r.h.s. endogenous variables
# x  = the complete set of instruments
# x1 = r.h.s. exogenous variables
# y1 = dichotomous l.h.s. variable

list y2 = eqrat bonus optval
list x = const ltass linsown linstown roe mktbk perfor \
     dealdum div dum97 dum98 dum99 dum00 no_emp no_subs \
     no_off ceo_age gap cfa
list x1 = const ltass linsown linstown roe mktbk perfor \
     dealdum div dum97 dum98 dum99 dum00
list y1 = d2

matrix X = { x }
matrix Y = { y2 }
matrix Y1 = { y1 }
matrix X1 = { x1 }
matrix Z = X1~Y

matrix b = invpd(X'*X)*X'*Y
matrix d = invpd(X'X)*X'Z

scalar kx = cols(X)
scalar ky = cols(Y)
scalar s = cols(Y)

loop foreach i y2
    ols $i x --quiet
    genr uhat$i = $uhat
    genr yhat$i = $yhat
endloop

matrix d = invpd(X'X)*X'Z

# step 2  RF probit
```

```
probit y1 x uhat* --quiet
genr J = $vcv
matrix alph = $coeff
matrix alpha = alph[1:kx]
matrix lam = alph[kx+1:kx+ky]
matrix Jinv=J[1:kx,1:kx]

# Step 3 2siv
probit y1 x1 uhat* yhat* --quiet
matrix beta = $coeff
matrix beta = beta[rows(beta)-ky+1:rows(beta)]
matrix rho = lam - beta

# step 4 v2*inv(x'x)
matrix rhoY=Y*rho
series ry =  rhoY
ols ry x --quiet
matrix v2 = $vcv

matrix omega = (v2+Jinv)

# Step 5
matrix cov = invpd(d'*invpd(omega)*d)
matrix se = sqrt(diag(cov))
matrix delt = cov*d'*invpd(omega)*alpha
print delt se
```

This code could be used as the basis for a more elegant gretl function that could be used to estimate this model. Basically, one just needs to load the data and replace the variable names to be used in the list statements. This version of the code illustrates just how easy it is to perform matrix computations in gretl in that the code mimics the steps listed in section 3.2.

One of the very useful properties of gretl is the way in which matrix computations and native gretl results can be intermixed. In this case, the usual probit mle can be estimated using native gretl routines and the resulting variance covariance matrix can be saved, converted to a matrix and used in subsequent computations. The –quiet option reduces the amount of output to a manageable level.

7 Conclusion

In this paper a simple gretl script is used to estimate the parameters of an dichotomous choice model that contains endogenous regressors. The routine is simple and yields the same results as the two-step option in the commercially available Stata 10 software.

The next step is to duplicate the maximum likelihood estimator, a considerably more challenging undertaking given the multitude of ways the mle can be computed. It should be noted that the only other commericial software that estimates this model via mle is Limdep; Limdep and Stata use different algorithms and yield different results.

Another possibility is to use the plug-in IVP estimator with Murphy-Topel standard errors. In very preliminary research Adkins [3] finds that this estimator compares favorably to AGLS and ML estimation in approximating the nominal size of 10% tests of parameter significance. Like the AGLS estimator, this should also be a relatively simple computation in gretl.

Bibliography

[1] Adkins, Lee C. [2008*a*], Small sample performance of instrumental variables probit estimators: A monte carlo investigation.

[2] Adkins, Lee C. [2008*b*], 'Small sample performance of instrumental variables probit estimators: A monte carlo investigation', Department of Economics, Oklahoma State University, Stillwater OK 74078. available at http://www.learneconometrics.com/pdf/JSM2008.pdf.

[3] Adkins, Lee C. [2009], A comparison of two-step and ml estimators of the instrumental variables probit estimators: A monte carlo investigation.

[4] Adkins, Lee C., David A. Carter and W. Gary Simpson [2007], 'Managerial incentives and the use of foreign-exchange derivatives by banks', *Journal of Financial Research* **15**, 399–413.

[5] Arendt, Jacob Nielsen and Anders Holm [2006], Probit models with binary endogenous regressors, Discussion Papers on Business and Economics 4/2006, Department of Business and Economics Faculty of Social Sciences University of Southern Denmark.

[6] Blundell, Richard W. and James L. Powell [2004], 'Endogeneity in semiparametric binary response models', *Review of Economic Studies* **71**, 655–679. available at http://ideas.repec.org/a/bla/restud/v71y2004ip655-679.html.

[7] Davidson, Russell and James G. MacKinnon [2004], *Econometric Theory and Methods*, Oxford University Press, Inc., New York.

[8] Greene, William H. [2007], *LIMDEP Version 9.0 Econometric Modeling Guide, Volume 1*, Econometrics Software, Inc., 15 Gloria Place, Plainview, NY.

[9] Iwata, Shigeru [2001], 'Recentered and rescaled instrumental variable estimation of tobit and probit models with errors in variables', *Econometric Reviews* **24**(3), 319–335.

[10] Kan, Kamhon and Chihwa Kao [2005], Simulation-based two-step estimation with endogenous regressors, Center for Policy Research Working Papers 76, Center for Policy Research, Maxwell School, Syracuse University. available at http://ideas.repec.org/p/max/cprwps/76.html.

[11] Murphy, Kevin M. and Robert H. Topel [1985], 'Estimation and inference in two-step econometric models', *Journal of Business and Economic Statistics* **3**(4), 370–379.

[12] Newey, Whitney [1987], 'Efficient estimation of limited dependent variable models with endogenous explanatory variables', *Journal of Econometrics* **36**, 231–250.

[13] Nicoletti, Cheti and Franco Peracchi [2001], Two-step estimation of binary response models with sample selection, Technical report, Faculty of Economics, Tor Vergata University, I-00133 Rome, Italy. Please do not quote.

[14] Rivers, D. and Q. H. Vuong [1988], 'Limited information estimators and exogeneity tests for simultaneous probit models', *Journal of Econometrics* **39**(3), 347–366.

[15] Shleifer, A. and R. W. Vishny [1986], 'Large shareholders and corporate control', *Journal of Political Economy* **94**, 461–488.

[16] Smith, Richard J. and Richard W. Blundell [1985], 'An exogeneity test for a simultaneous equation tobit model with an application to labor supply', *Econometrica* **54**(3), 679–685.

[17] Tufano, P. [1996], 'Who manages risk? an empirical examination of risk management practices in the gold mining industry', *Journal of Finance* **51**, 1097–1137.

[18] Whidbee, D. A. and M. Wohar [1999], 'Derivative activities and managerial incentives in the banking industry', *Journal of Corporate Finance* **5**, 251–276.

[19] Yatchew, Adonis and Zvi Griliches [1985], 'Specification error in probit models', *The Review of Economics and Statistics* **67**(1), 134–139.

Automatic Procedure of Building Congruent Dynamic Model in Gretl

Marcin Błażejowski, Paweł Kufel, and Tadeusz Kufel

Toruń School of Banking,
Młodzieżowa 31a, 87-100 Toruń, Poland
marcin.blazejowski@wsb.torun.pl
Nicolaus Copernicus University,
Department of Econometrics and Statistics,
Gagarina 13a, 87-100 Toruń, Poland
qfel@mat.uni.torun.pl,
tadeusz.kufel@uni.torun.pl

Abstract. In the last years we can observe intensive development of automatic model selection procedures. Best known are PcGets and RETINA. Such intensive work encourage to work on a new procedures. The concept of Congruent Modelling, formulated by Prof. Zygmunt Zieliński, is a very good framework for such development, including programming work, as well as many theoretical considerations. In the paper we present our concept of algorithm for automatic congruent modelling procedure and propose it's implementation in Gretl.

Key words: congruent dynamic modelling, automatic model selection, forecasting, PcGets, RETINA

1 Introduction

In article [3] there is a very interesting dialog between Prof. Granger and Prof. Hendry about PcGets – automatic model selection procedure described in [5]. Prof. Granger had formulated 20 questions concerning specification of GUM, simplification the GUM, testing economic theories, policy applications, nonstationarity, nonlinearity, multiple equation models and forecasting and asked Prof. Hendry how PcGets – automatic model selection procedure – handles with it.

That dialog shows how automatic model selection tools are important in contemporary econometrics, but also shows how difficult this area is. There are two major automatic model selection procedures – PcGets[1] and RETINA, described in [9]. The aim of this paper is formulation of algorithm for automatic

[1] Since PcGive 12 and OxMetrics 5, PcGets is no longer available and current automatic model selection procedure is called Autometrics (see [2]). In this article we still use the former name, but all our considerations on PcGets refer to Autometrics as well.

model selection procedure based on congruent modelling approach and – as the next step – implementation that algorithm in Gretl[2].

1.1 Idea of congruent modeling

The congruent modeling refers to building dynamic econometric models and was presented by Prof. Zygmunt Zieliński from Nicolaus Copernicus University from Toruń in 1984.

Many assumptions underlay the formulating of initial model specification. Some approaches refere to causal relationships, the other ones – to the internal structure of processes of interest with omission of causality, and others take both into account. The concept of congruent modeling, in Zieliński sense, refers to both approaches – casual relationship and internal structure of given processes[3].

A model is congruent, according to Zieliński, if the harmonic structure of dependent process Y_t is the same as the joint harmonic structure of explanatory processes X_{it} $(i = 1, 2, \ldots, k)$ and the residual process, which is independent of explanatory processes. This means that the variability of left side of model – (Y_t) must be explained by the variability of right side of model – (X_{it}). It is obvious that the model built for processes having white noise properties is always congruent:

$$\varepsilon_{yt} = \sum_{i=1}^{k} \rho_i \varepsilon_{x_i t} + \varepsilon_t, \tag{1}$$

where $\varepsilon_{yt}, \varepsilon_{x_i t}$ and ε_t are white noises. Model (1) is congruent because harmonic structure of both sides of equation are equal or, in other words, the processes spectra are parallel to the frequency axis.

Let Y_t and X_{it} $(i = 1, 2, \ldots, k)$ denote the endogenous process and explanatory processes respectively with the internal structure of:

– models describing non-stationary components:

$$\begin{aligned} Y_t &= P_{yt} + S_{yt} + \eta_{yt}, \\ X_{it} &= P_{x_i t} + S_{x_i t} + \eta_{x_i t}, \end{aligned} \tag{2}$$

where $P_{yt}, P_{x_i t}$ are polynomial functions of variable t, $S_{yt}, S_{x_i t}$ denote seasonal component with constant or changing in time amplitude of fluctuations and $\eta_{yt}, \eta_{x_i t}$ are stationary autoregressive processes for respective variables, and

[2] Any information about Gretl (Gnu Regression, Econometrics and Time-series), can be found in [1].

[3] More on congruent dynamic modelling one can find in: [13], [16], [14], [7], [11], [12].

– autoregressive processes:

$$B(u)\eta_{yt} = \varepsilon_{yt}, \tag{3}$$
$$A_i(u)\eta_{x_it} = \varepsilon_{x_it}, \tag{4}$$

where $B(u), A_i(u)$ denote stationary autoregressive back shift operators for which all roots of equations $|B(u)| = 0$ and $|A_i(u)| = 0$ lie outside the unit root circle and $\varepsilon_{yt}, \varepsilon_{x_it}$ are white noises for respective processes.

Information about internal structure of Y_t and X_{it} processes enable to build the congruent dynamic econometric model by substituting ε_{yt} and ε_{x_it} in model (1) from models (3) and next for autoregressive processes in models (2). After transformations the congruent general dynamic econometric model is as follows:

$$B(u)Y_t = \sum_{i=1}^{k} A_i^*(u)X_{it} + P_t + S_t + \varepsilon_t, \tag{5}$$

where $B(u), A_i^*(u)$ are autoregressive back shift operators, P_t is polynomial function of variable t, S_t denotes seasonal component with constant or changing in time amplitude of fluctuations and ε_t is white noise. The white noise ε_t in model (5) has the same properties as white noise ε_t in model (1). Whole information of internal structure of all processes is taken into consideration. The variability of endogenous process Y_t is explained by variability of exogenous processes $X_{it}, (i = 1, \ldots, k)$.

Described concept of building dynamic econometric model shows the necessity of including information about internal structure of given processes at the model specification stage.

1.2 Linear congruent model for intergrated processes

Let[4] assume that endogenous Y_t and exogenous X_{it} are intergrated processes with zero mean of order, respectively, $d_y \geq 1$ and $d_{x_i} \geq 1$. It means, that:

$$\begin{aligned} Y_t^* &= (1-u)^{d_y} Y_t = \Delta^{d_y} Y_t \,, \\ X_{it}^* &= (1-u)^{d_{x_i}} X_{it} = \Delta^{d_{x_i}} X_{it} \,, \end{aligned} \tag{6}$$

are covariance stationary processes with zero mean, where u is such back shift operator, that $u^s Z_t = Z_{t-s}$. Processes Y_t^* and X_{it}^* can be expressed in AR notation:

$$\begin{aligned} B(u)Y_t^* &= \varepsilon_{yt} \,, \\ A_i(u)X_{it}^* &= \varepsilon_{x_it} \,, \end{aligned} \tag{7}$$

[4] This paragraph is based on [15] and [14].

where ε_{yt} and $\varepsilon_{x_i t}$ are white noises with zero mean, $B(u)$ and $A_i(u)$ are stationary autoregressive operators of order, respectively, p_y and q_{x_i}.

Operators:

$$B^*(u) = B(u)(1-u)^{d_y} = 1 - \sum_{s=1}^{p_y+d_y} \alpha_s^* u^s = 1 - B_1^*(u)\,,$$
$$A_i^*(u) = A_i(u)(1-u)^{d_{x_i}} = 1 - \sum_{s=1}^{p_{x_i}+d_{x_i}} \gamma_s^* u^s = 1 - A_{i1}^*(u)\,, \qquad (8)$$

are nonstationary autoregressive operators which satisfy condition, that d_y or d_{x_i} roots of $B^*(z) = 0$ or $A_i^*(z) = 0$ lie on a unit circle. Taking (8) into account, nonstationary processes Y_t and X_{it} can be expressed as:

$$Y_t = B_1^*(u)Y_t + \varepsilon_{yt}\,,$$
$$X_{it} = A_{i1}^*(u)X_{it} + \varepsilon_{x_i t}\,. \qquad (9)$$

Linear congruent model (1), describing relationship between Y_t and X_{it}, can be write down in the following form:

$$Y_t = \sum_{i=1}^{k} \rho_i X_{it} - \sum_{i=1}^{k} \rho_i A_{i1}^*(u)X_{it} + B_1^* Y_t + \varepsilon_t\,, \qquad (10)$$

Coefficient ρ in (10) is the same as in (1). Orders of autoregressions in model (10) are extended of orders of integration, which means, that lags of Y_t are equal to $p_y + d_y$ and lags of X_{it} are equal to $p_{x_i} + d_{x_i}$.

2 General algorithm for automatic building congruent dynamic model

This is a very general algorithm for building congruent dynamic model, where only main stages are described without talking over any specific solutions, internal variables, used external functions and so on. This algorithm shows only a general idea of our procedure and it's compatibility with congruent dynamic econometric modelling procedure in Zieliński sense.

1. Getting outgoing data, setting following internal variables:
   (a) Getting model variables: endogenous Y, list of explanatory X, list of deterministic (dummy) variables.
   (b) Getting range of the sample and setting minimal degrees of freedom *dfmin* for starting general congruent model:

   i. if $n < 200$, then $dfmin = round(0.1 \times n)$,
   ii. if $n \geq 200$, then $dfmin = 20$.
   (c) Checking the frequency of time-series and setting:
   i. deterministic cycle for consideration in pt. 2,
   ii. maximum order *pmax* for autoregressive models in pt. 2c,
2. Analysis the internal structure of given processes:
   (a) Checking, whether given processes have deterministic components.
   (b) Checking, whether error terms after subtraction of deterministic components are integrated.
   (c) Setting orders of autoregression for given processes, starting from a maximum order of *pmax*, after subtraction of deterministic components and differentiation if there was an integration.
3. Building starting specification of general unrestricted congruent model:
   (a) Checking the degrees of freedom of starting general congruent model *dfstart*, taking into account all possible variables (lagged Y, current and lagged X, trend and/or cycle, deterministic variables):
   i. if $dfstart < dfmin$, then maximum order of autoregressive model(s), specified in pt. 2c, is decreased by 1,
   ii. if $dfstart \geq dfmin$, then the starting general model is stored in Gretl session.
4. Building congruent empirical model (specific):
   (a) Specified in pt. 3 starting congruent general model is reduced according to *a posterior* procedure of variable selection with use of *t* statistics.
   (b) Congruent empirical model is stored in Gretl session.

At **stage 1**, we just import dataset for model: endogenous Y, list of "normal" explanatory X and some deterministic (dummy) variables, which indicate some special moments, i.e. free days. For next stages, we check dataset structure.

At **stage 2** we perform the analysis of internal structure of explained process Y_t and explanatory processes X_{it}. We are looking for deterministic components: time trend and periodicity. We assume, that we only check presence of linear trend and "typical" periodicity for given data: seasonality for monthly or quarterly time-series, 1 year (52 weeks) cycle for weekly time-series and 1 week cycle for daily data. After that we subtract deterministic components from analyzed processes and perform ADF test. If there is an integration, we difference series and examine order of autoregression of all time-series.

At **stage 3** we formulate starting congruent general model and check, if we have sufficient degrees of freedom for running OLS estimation. This is crucial, that we want at least $0.1 \times n$ degrees of freedom or, if series have more then 200 observations, at least 20. If we can't meet this condition, maximum order

of autoregressive models is decreased by 1, so we earn at least 2 degrees of freedom at each reduction.

At **stage 4** we perform OLS estimation of coefficients of starting general congruent model with strategy of model reduction based on *a posterior* elimination with *t* statistics as a criterion. Empirical congruent model build on above algorithm always has white noise error terms.

3 Comparison of Congruent Modelling algorithm vs. PcGets vs. RETINA

In this section there are shown some similarities and differences beetwen algorithm based on congruent dynamic modelling theory and approaches in PcGets and RETINA. PcGets and RETINA comparison one can find in [10] and tables (1)-(9) are based on it. There are just delivered suitable information about our approach, so comparison of that three automatic model selection procedures is now possible and easy. Comparison of General-to-Specific vs. Congruent Modelling one can find in [8].

Table 1. Goals

PcGets	RETINA	Congruent Modelling
1. Select a parsimonious undominated representation of an overly general initial model, the general unrestricted model (GUM).	1. Identify a parsimonious set of transformed attributes likely to be relevant for predicting out-of-sample.	1. Congruent general model is reduced parsimonious congruent model in Zieliński sense, which means error term of white noise properties.
2. Best model fit within sample.		2. Very good behavior in prediction out of the sample.
3. Congruent with theory.		3. Congruent with theory.

In congruent modelling we believe, that DGP is nested in the starting general model and reduction irrelevant variables can discover it. This is very similar to general-to-specific approach, but the starting general model is formulated in different ways. Starting specification is based on theory and extended of information about internal structure of given processes, including deterministic components, integration and autoregression. Autoregressive models can have different starting order, which is the biggest difference with general-to-specific approach.

In congruent modelling we start with congruent general model and step by step this model is reduced to congruent empirical model with white noise error

Table 2. Strategy

PcGets	RETINA	Congruent Modelling
1. General to specific. 2. Formulate a GUM and reduce it to a parsimonious model using residual tests and hypothesis testing on coefficients.	1. Specific to general: Start from a model with a single transform. Add additional transforms only if they contribute to out-of-sample forecast ability. 2. Flexible and parsimonious model. 3. Selective search of transforms. 4. Control for collinearity.	1. Congruent general model is reduced to congruent empirical model. 2. Elimination insignificant variables using *a posterior* procedure based on *t* statistics. 3. Empirical model is parsimonious.

term. As a strategy we use *t* statistics which has sufficient power to discover DGP (see numerical experiment).

Table 3. Base Model

PcGets	RETINA	Congruent Modelling
1. GUM: specified by the researcher, usually based on theory. May use transforms of the original variables.	1. Based on original inputs and transforms, automatically selected from the first subsample by cross-validation in the second, controlling for collinearity.	1. Specification of starting general congruent is based on a theory and extended of information of internal structure of all included processes. 2. Congruent model may use transformed processes.

Base model is formulated on two basis: theory, which gives us causal relationships between variables and on the internal structure of all processes. This guarantee, that hole variability of Y_t and all X_{it} processes is included, so model is congruent (error term has white noise properties).

Congruent modelling assumes linear in parameters model, but variables can be log-transformed. Model can be nonlinear in variables, so congruent modelling gives maximum flexibility.

Starting congruent general model is unrestricted and, because of specification, based on the internal structure of all processes, overparametrizied, but step by step model is being reduced with use of *t* statistics and *a posterior* procedure. Final (empirical) congruent model is parsimonious.

Table 4. Flexibility

PcGets	RETINA	Congruent Modelling
1. The GUM determines maximum flexibility. May include transforms of the original variables.	1. The permitted transformations of the inputs determine maximum flexibility. 2. The actual flexibility of the candidate model is chosen by the procedure.	1. Congruent general model is unrestricted, so it gives maximum flexibility.

Table 5. Selective/Systematic Search

PcGets	RETINA	Congruent Modelling
1. Starting from the GUM, performs a systematic search using multiple reduction paths. 2. Using diagnostics, checks the validity of each reduction until terminal selection. 3. When all paths are explored, repeatedly tests models against their union until a unique final model is obtained.	1. Uses a selective search to avoid the heavy task of evaluating all 2^m possible models and of applying some form of model selection. 2. A saliency feature of the transforms, such as the correlation with the dependent variable, is used to construct a natural order of the transforms in which they are considered. 3. Only a number of candidate models of order proportional to m is considered.	1. Starting from congruent general model and reduce it by eliminating irrelevant variables with *t* statistics.

Table 6. Colinearity

PcGets	RETINA	Congruent Modelling
1. Seeks to formulate the GUM, in search for a relatively orthogonal specification. 2. A quick modeler option is available in PcGets for nonexpert users.	1. Controls for collinearity by adding an additional transform to the candidate list only if the collinearity is below a certain (user defined) threshold.	1. Colinearity is controlled by GRETL.

Table 7. Explanatory Variables

PcGets	RETINA	Congruent Modelling
1. Original variables and transformations specified in the GUM.	1. Original variables and nonlinear transformations allowed for by the procedure.	1. Original variables and handmade transformations specified at the beginning of procedure.

Table 8. Linearity

PcGets	RETINA	Congruent Modelling
1. Linear or nonlinear in the parameters, as specified by the GUM. 2. Linear or nonlinear in the underlying variables, as specified by the GUM.	1. Linear in the parameters. 2. Linear or nonlinear in the underlying variables.	1. Linear in the parameters. 2. Linear or nonlinear in the underlying variables.

Congruent model assumes linearity in parameters, so computation is simple (we use OLS estimation). Transformations of variables are allowed, so starting general model can be nonlinear in the underlying variables.

Table 9. Types of Data Applicable So Far

PcGets	RETINA	Congruent Modelling
1. Time series or cross-section.	1. Mainly cross-section at present (no obstacles to its application in a time series context).	1. Time-series.

Congruent modelling is applicable to time-series and cross-section data as well, but our automatic procedure assumes time-series only. Congruent modelling approach is also applicable to multiple equation systems, including simultaneous equation models, but automatic procedure for it would be very complicated (but not impossible).

4 Numerical experiment

To introduce the efficiency of model selection using the congruent modeling postulate there is numerical experiment presented. The experiment is based on Monte Carlo simulations. The scenario of experiment is summarized in table (10).

We assume situation (which is actually very common real case), that we do not have any observations of Z_t process, which is the component of DGP process. Because of autoregressive internal structure of *X* and *Z* processes ($X_t, Z_t \sim AR(1)$), specification of starting congruent general model in table (10) was based on internal autoregressive structure of $X_t \sim AR(1)$ and $Y_t \sim AR(2)$ or $AR(3)$, as a result of combination of two $AR(1)$ processes (see [4]).

For both scenarios following steps were realized:

Table 10. Experimental design

DGP	
$Y_t = 3X_t + 3Z_t + \varepsilon_{yt}$	$\varepsilon_{yt} \sim IN(0, \sigma_y^2)$
$X_t = \beta_x X_{t-1} + \varepsilon_{xt}$	$\varepsilon_{xt} \sim IID(0, 1)$
$Z_t = \beta_z Z_{t-1} + \varepsilon_{zt}$	$\varepsilon_{zt} \sim IID(0, 1)$
$\varepsilon_{xt} = \rho\varepsilon_{zt}$	$t = 1, 2 \ldots, n$
Congruent General Model	
$Y_t = \alpha_0 + \alpha_1 X_t + \alpha_2 X_{t-1} + \alpha_3 Y_{t-1} + \alpha_4 Y_{t-2} + \alpha_5 Y_{t-3} + \varepsilon_t$	$\varepsilon_t \sim IID$
Experiment A $(\alpha = 0.01)$ – table (11)	
DGP:	
$n = \{300, 120, 60, 20\}$	
$\rho = \{0.0, 0.2, 0.4, 0.6, 0.8\}$, $\sigma_y = \{1, 3\}$	
$\beta_x = \{0.6, 0.8, 0.95\}$, $\beta_z = \{0.6, 0.8, 0.95\}$	
Experiment B $(\alpha = 0.05)$ – table (12)	
DGP:	
$n = \{300, 120, 60, 20\}$	
$\rho = \{0.0, 0.2, 0.4, 0.6, 0.8\}$, $\sigma_y = \{1, 3\}$	
$\beta_x = \{0.6, 0.8, 0.95\}$, $\beta_z = \{0.6, 0.8, 0.95\}$	

1. Coefficients of the starting congruent general model, specified according to congruence postulate and formulated in table (10), was estimated by OLS method.
2. Elimination of insignificant processes was based on *t*-Student statistics and realized according to *a posterior* procedure at significance level of $\alpha = \{0.01, 0.05\}$.
3. Encompassing *J* test was run verifying the null hypothesis that the empirical model is special case of DGP. The number of not rejected null hypothesis was compared.

Tables (11) and (12) present percentage of non rejection the null hypothesis that the empirical congruent model is a special case of DGP. Results of *J* test show, that for samples of $n = \{300, 120\}$ all cases of empirical congruent models, which were build without relevant process Z_t (one of the components of DGP process), was a special case of the data generating process. For samples $n = 60$ and noise $\varepsilon \sim N(0, 1)$ percentage of discover of DGP was 99%–100% and for noise $\varepsilon \sim N(0, 9)$ it was 87%–100%. For samples $n = 20$ percentage of discovering DGP was much lower and for noise $\varepsilon \sim N(0, 1)$ it was 70%–95% and for noise $\varepsilon \sim N(0, 9)$ it was 41%–88%.

So the conclusion is, that *a posterior* elimination procedure based on *t* statistics has sufficient power and even for small samples, percentage of discovering DGP is still relatively high and has a value of about 70%–80%.

Table 11. Percentage of not rejecting the null hypothesis assuming that congruent model is special case of DGP for $\alpha = 0.01$

			β_x=0,6			β_x=0,8			β_x=0,95		
n	ε_t	ρ	β_z=0,6	β_z=0,8	β_z=0,95	β_z=0,6	β_z=0,8	β_z=0,95	β_z=0,6	β_z=0,8	β_z=0,95
n=300	$\varepsilon_t(0,1)$	ρ =0,0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$\varepsilon_t(0,3)$	ρ =0,0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
n=120	$\varepsilon_t(0,1)$	ρ =0,0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$\varepsilon_t(0,3)$	ρ =0,0	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
n=60	$\varepsilon_t(0,1)$	ρ =0,0	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,2	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,4	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,6	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,8	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$\varepsilon_t(0,3)$	ρ =0,0	0.87	0.98	1.00	0.98	1.00	1.00	1.00	1.00	1.00
		ρ =0,2	0.92	0.98	1.00	0.99	1.00	1.00	0.99	1.00	1.00
		ρ =0,4	0.92	0.98	1.00	0.99	1.00	1.00	1.00	1.00	1.00
		ρ =0,6	0.94	0.99	1.00	0.98	1.00	1.00	1.00	1.00	1.00
		ρ =0,8	0.95	0.99	1.00	0.98	1.00	1.00	1.00	1.00	1.00
n=20	$\varepsilon_t(0,1)$	ρ =0,0	0.85	0.85	0.92	0.83	0.89	0.93	0.88	0.91	0.96
		ρ =0,2	0.76	0.86	0.91	0.81	0.88	0.94	0.87	0.91	0.95
		ρ =0,4	0.75	0.83	0.87	0.82	0.88	0.92	0.85	0.92	0.95
		ρ =0,6	0.73	0.83	0.86	0.84	0.88	0.92	0.87	0.90	0.94
		ρ =0,8	0.70	0.80	0.83	0.82	0.88	0.90	0.84	0.90	0.93
	$\varepsilon_t(0,3)$	ρ =0,0	0.61	0.68	0.75	0.73	0.81	0.84	0.81	0.87	0.88
		ρ =0,2	0.54	0.66	0.72	0.67	0.78	0.79	0.77	0.82	0.85
		ρ =0,4	0.51	0.59	0.70	0.63	0.72	0.79	0.72	0.80	0.82
		ρ =0,6	0.47	0.56	0.67	0.59	0.67	0.75	0.72	0.80	0.79
		ρ =0.8	0.41	0.51	0.56	0.54	0.60	0.68	0.66	0.71	0.73

Table 12. Percentage of not rejecting the null hypothesis assuming that congruent model is special case of DGP for $\alpha = 0.05$

			β_x=0,6			β_x=0,8			β_x=0,95		
n	ε_t	ρ	β_z=0,6	β_z=0,8	β_z=0,95	β_z=0,6	β_z=0,8	β_z=0,95	β_z=0,6	β_z=0,8	β_z=0,95
n=300	$\varepsilon_t(0,1)$	ρ =0,0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$\varepsilon_t(0,3)$	ρ=0,0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
n=120	$\varepsilon_t(0,1)$	ρ=0,0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$\varepsilon_t(0,3)$	ρ=0,0	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ =0,8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
n=60	$\varepsilon_t(0,1)$	ρ=0,0	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ=0,2	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ=0,4	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ=0,6	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ρ=0,8	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$\varepsilon_t(0,3)$	ρ=0,0	0.89	0.98	1.00	0.98	0.99	1.00	1.00	1.00	1.00
		ρ=0,2	0.91	0.98	1.00	0.98	1.00	1.00	0.99	1.00	1.00
		ρ=0,4	0.93	0.99	1.00	0.98	1.00	1.00	1.00	1.00	1.00
		ρ=0,6	0.93	0.99	1.00	0.98	1.00	1.00	0.99	1.00	1.00
		ρ=0,8	0.95	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00
n=20	$\varepsilon_t(0,1)$	ρ=0,0	0.73	0.85	0.91	0.80	0.89	0.92	0.84	0.92	0.94
		ρ=0,2	0.73	0.81	0.87	0.78	0.89	0.93	0.88	0.91	0.91
		ρ=0,4	0.70	0.82	0.85	0.79	0.86	0.90	0.84	0.90	0.91
		ρ=0,6	0.72	0.77	0.84	0.77	0.88	0.89	0.81	0.90	0.92
		ρ=0,8	0.70	0.76	0.81	0.80	0.87	0.88	0.82	0.87	0.93
	$\varepsilon_t(0,3)$	ρ=0,0	0.53	0.61	0.70	0.62	0.72	0.77	0.76	0.79	0.82
		ρ=0,2	0.50	0.57	0.68	0.61	0.69	0.77	0.70	0.77	0.83
		ρ=0,4	0.51	0.57	0.65	0.57	0.69	0.77	0.68	0.75	0.83
		ρ=0,6	0.45	0.55	0.65	0.59	0.67	0.72	0.66	0.76	0.79
		ρ=0.8	0.41	0.47	0.56	0.52	0.55	0.64	0.64	0.70	0.70

5 Summary

In the paper we discussed power of the congruent modelling concept and it is intrinsic features for being the base for automatic model selection procedure. Although we formulated full algorithm for such procedure, our considerations was theoretical. So the next stage of our work will be implementation this algorithm in Gretl. Our automatic procedure will:

1. Investigate internal trend-seasonal-autoregressive structure for all processes.
2. Formulate congruent general model on the basis of initial list of explanatory extended of it's internal components.
3. Run OLS estimation and eliminate insignificant variables according to *a posterior* procedure based on t statistics.
4. Store empirical congruent model in Gretl session.

Bibliography

[1] Cottrell, A., Lucchetti, R.: Gnu Regression, Econometrics and Time-series, `http://gretl.sourceforge.net`

[2] Doornik, J. A., Hendry, D. F.: Empirical Econometric Modelling – PcGive 12. Timberlake Consultants Press, London (2007)

[3] Granger, C. W. J., Hendry, D. F.: A dialogue concerning a new instrument for econometric modeling. Econometric Theory, vol. 21, pp. 278–297, (2005)

[4] Granger, C. W. J., Morris, M. J.: Time Series Modelling and Interpretation. Journal of the Royal Statistical Society. Series A (General). 139(2), 246–257 (1976)

[5] Hendry, D. F., Krolzig, H.-M.: Automatic Econometric Model Selection. Timberlake Consultants Press, London (2001)

[6] Krolzig, H.-M., Hendry, D. F.: Computer automation of general-to-specific model selection procedures. Journal of Economic Dynamics and Control. 25, 831–866 (2001)

[7] Kufel, T, Piłatowska, M.: The White Noise Properties in the Dynamic Congruent Modelling. In: MACROMODELS '90, pp. 163–178. Łódź University Press, Łódź (1997)

[8] Kufel, T.: General-to-Specific Modelling vs. Congruent Modelling in PcGets. In: Zieliński, Z. (ed.) Dynamic Econometric Models, vol. 6, pp. 83–92. Nicolaus Copernicus University Press, Toruń (2004)

[9] Perez-Amaral, T., Gallo, G. M., White, H.: Flexible Tool for Model Building: the Relevant Transformation of the Inputs Network Approach (RETINA). Technical report, Universidad Complutense de Madrid, Facultad de Ciencias Económicas y Empresariales (2003)

[10] Perez-Amaral, T., Gallo, G. M., White, H.: A Comparison of Complementary Automatic Modeling Methods: RETINA and PcGets. Econometric Theory, Vol. 21, No. 1, pp. 262–277, (2005)

[11] Piłatowska, M.: Realization of the Congruence Postulate as a Method of Avoiding the Effect of a Spurious Relationship. In: Zieliński, Z. (ed.) Dynamic Econometric Models, vol. 6, pp. 117–126. Nicolaus Copernicus University Press, Toruń (2004)

[12] Piłatowska, M.: The Econometric Models Satisfying the Congruence Postulate – an Overview. In: Zieliński, Z. (ed.) Dynamic Econometric Models, vol. 8, pp. 53–60. Nicolaus Copernicus University Press, Toruń (2008)

[13] Zieliński, Z.: Dynamic Econometric Linear Models as a Tool of Description and Analysis of Causal Relationships in Economics. In: Zieliński, Z.

(ed.) Some Aspects of the Dynamic Eeconometric Modelling, pp. 7–58. Nicolaus Copernicus University Press, Toruń (1993)

[14] Zieliński, Z.: Linear Congruent Models Describing Relationships for Integrated Economic Processes. In: Zieliński, Z. (ed.) Dynamic Econometric Models, vol. 1, pp. 7–20. Nicolaus Copernicus University Press, Toruń (1994)

[15] Zieliński, Z.: Liniowe modele zgodne opisujace zależności sumacyjnych (zintegrowanych) procesów ekonomicznych. In: Zeliaś, A. (ed.) Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych, pp. 77–87. AE Kraków, Kraków (1995)

[16] Zieliński, Z., Kufel, T.: Specification of Dynamic Properties of the Econometric Models in the Light of Congruent Models Concept. In: MACROMODELS '86, pp. 25–52. Łódź University Press, Łódź (1995)

Instrumental Variable Interval Regression

Giulia Bettin[1] and Riccardo (Jack) Lucchetti[2]

[1] HWWI - Hamburg Institute for International Economics - Hamburg, Germany
bettin@hwwi.org
[2] Dipartimento di Economia - Universita Politecnica delle Marche - Ancona, Italy
r.lucchetti@univpm.it

Abstract. In this paper, we introduce a maximum-likelihood estimator for grouped data with endogenous regressors and briefly analyse its properties. An example application to migrants' remittances is included, which shows that endogeneity effects, so far neglected by the applied literature, are substantial.

1 Introduction

The estimation of interval models by maximum likelihood, introduced by [16], is nowadays relatively straightforward and has been applied in a number of contexts, most notably in willingness-to-pay double bound models (for a recent example, see [15]). The data generating process is assumed to be

$$y_i^* = x_i'\beta + \epsilon_i \tag{1}$$

where y_i^* is unobservable *per se*; what is observed are the limits of an interval that contains it, that is

$$m_i \leq y_i^* \leq M_i$$

where the interval may be left- or right-unbounded. Once a distributional hypothesis for ϵ_i is made, estimation becomes a simple application of maximum likelihood techniques. Under normality, the log-likelihood for one observation is

$$\ell_i(\beta, \sigma) = \ln P(m_i < y_i^* \leq M_i) = \ln \left[\Phi\left(\frac{M_i - x_i'\beta}{\sigma}\right) - \Phi\left(\frac{m_i - x_i'\beta}{\sigma}\right) \right] \tag{2}$$

and the total log-likelihood can be maximised by standard numerical methods, which are, in most cases, very effective. The above procedure is implemented natively in several econometric packages, among which Gretl, Limdep, Stata and TSP.

However, the extension of this model to the case of endogenous regressors seems to be absent from the literature. To consider this case, equation (1) can be

generalised to

$$y_i^* = Y_i'\beta + X_i'\gamma + \epsilon_i \tag{3}$$

$$Y_i = X_i\Pi_1 + Z_i\Pi_2 + u_i = W_i\Pi + u_i \tag{4}$$

$$\begin{bmatrix} \epsilon_i \\ u_i \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma^2 & \theta' \\ \theta & \Sigma \end{bmatrix}\right) \tag{5}$$

where θ, the covariance between ϵ_i and u_i may be nonzero. In this case, the vector of m explanatory variables Y_i becomes endogenous and ordinary interval regression does not provide consistent estimates of β and γ.

2 Estimation methods

2.1 Limited-information Maximum Likelihood

The estimation problem can be tackled by maximum likelihood. If y_i^* were observable, the log-likelihood for one observation could be written as follows

$$\ell_i(\psi) = \ln f(\epsilon_i, u_i; \psi) = \ln\left[f(\epsilon_i|u_i; \psi)\right] + \ln f(u_i; \psi) = \ell_i^a + \ell_i^b \tag{6}$$

where ψ is a vector containing all the parameters, possibly transformed via an invertible mapping. Since y_i^* is imperfectly observed, however, the above has to be modified as

$$\ell_i^a(\psi) = \ln P(m_i < y_i^* \leq M_i|u_i) \tag{7}$$

Thanks to the assumed joint normality, the distribution of $(\epsilon_i|u_i)$ is

$$\epsilon_i|u_i \sim N\left(u_i'\lambda, \tilde{\sigma}^2\right)$$

where $\lambda = \Sigma^{-1}\theta$ and $\tilde{\sigma}^2 = \sigma^2 - \theta'\Sigma^{-1}\theta$. Hence,

$$\ell_i^a = \ln P(m_i < y_i^* \leq M_i|u_i) = \ln\left[\Phi\left(\frac{M_i - \hat{y}_i}{\tilde{\sigma}}\right) - \Phi\left(\frac{m_i - \hat{y}_i}{\tilde{\sigma}}\right)\right]$$

where $\hat{y}_i = Y_i'\beta + X_i'\gamma + u_i'\lambda$, and ℓ_i^b is just an ordinary normal log-likelihood:

$$\ell_i^b = \ln f(u_i; \psi) = -1/2\left[m\ln(2\pi) + \ln|\Sigma| + (Y_i - W_i\Pi)'\Sigma^{-1}(Y_i - W_i\Pi)\right] \tag{8}$$

Of course, we assume that the instruments Z_i satisfy the order and rank identification conditions.

In order to guarantee that σ is positive during the numerical search, what is actually fed to the log-likelihood function is its logarithm. For similar reasons, the unconstrained parameters on which the log-likelihood function is based are

not the elements of Σ itself, but rather those of the Cholesky factorisation of its inverse. In practice, ℓ_i^b, the second component of the log-likelihood, is computed as

$$\ell_i^b = \text{const} + \ln |C| - \frac{\omega_i' \omega_i}{2}$$

where C is the Cholesky factorisation of Σ^{-1} and $\omega_i = C'(Y_i - \Pi' W_i)$. This produces faster and more accurate computation than evaluating (8) directly for two reasons: first, a matrix inversion is avoided; moreover, the determinant of C (which is by construction $|\Sigma|^{-1/2}$) is trivial to compute since C is triangular, via

$$-0.5 \ln |\Sigma| = \sum_{i=1}^{m} \ln C_{ii}.$$

The computational gain is negligible (arguably null) when $m = 1$, but may become substantial for $m > 1$; in fact, casual experimenting show non-negligible improvements even for $m = 2$.

A recent paper by [8] advocates the usage of the EM algorithm for dealing with numerical problems in a closely related case (the ordered probit model), but we found it unnecessary in our case.

The ML setup also enables us to build two hypothesis tests which are likely to be of interest: the first one is an exogeneity test, which is constructed by testing for $\lambda = 0$ by means of a Wald test. In addition, a LR test for overidentifying restrictions may be computed via the difference ℓ^a and that for an interval regression of (m_i, M_i) on W_i and $\hat{u}_i$, which would be the unrestricted log-likelihood.

2.2 Alternative estimators

In certain cases, it may be worthwhile to consider alternative estimators than ML. Two are briefly considered here, although no serious effort is made to analyse them in detail; both belong to the two-step category of estimators. As such, they may suffer from the typical shortcoming of two-step estimators: inefficiency and a cumbersome-to-compute covariance matrix[3]. Hence, we only sketch briefly the possibility for alternative estimators; proper analysis of their properties is left as a future project.

Both estimators depend on the availability of an ordinary interval regression routine. One possibility is:

1. perform first-stage OLS of Y_i on W_i and collect the residuals $\hat{u}_i$
2. perform an interval regression on Y_i, X_i and $\hat{u}_i$

[3] The obligatory reference here is [12], but see also [17], chapter 12.

This estimator ought to be consistent[4]. This estimator is very easy to compute if an interval regression routine is available: as a consequence, it was a natural choice for initialising our ML algorithm.

Another two-step estimator may be obtained by considering that, given an interval regression of the form (1), it is easy to form an unbiased estimator of y_i^* from

$$E(y_i^*|x_i, m_i, M_i) = x_i'\beta + E(\epsilon_i|x_i, m_i, M_i) = x_i'\beta + \sigma \frac{\varphi\left(\frac{m_i - x_i'\beta}{\sigma}\right) - \varphi\left(\frac{M_i - x_i'\beta}{\sigma}\right)}{\Phi\left(\frac{M_i - x_i'\beta}{\sigma}\right) - \Phi\left(\frac{m_i - x_i'\beta}{\sigma}\right)}$$

and substituting unknown parameters with their estimates to get the estimate $\hat{y}_i$. The procedure is the following:

1. do an interval regression of (m_i, M_i) on W_i; that is, estimate the unrestricted reduced form of equation (3).
2. compute $\hat{y}_i$, an unbiased estimate if y_i^*; By construction,

   $$v_i \equiv y_i^* - \hat{y}_i$$

   will have the property $E(v_i|W_i) = 0$.
3. do TSLS using $\hat{y}_i$ as the dependent variable; this should be valid since (from equation (3)

   $$\hat{y}_i = Y_i'\beta + X_i'\gamma + (\epsilon_i - v_i)$$

   and the composite error term $(\epsilon_i - v_i)$ is incorrelated with the instruments W_i (although it will be heteroskedastic by construction).

Again, we conjecture that this estimator should also be consistent, but like the other one, it would be inefficient and the estimation of the parameters' covariance matrix would need a two-step adjustment.

3 Why bother?

From the viewpoint of an applied economist, the ML method outlined above may seem overkill. After all, how much inaccuracy do we introduce in the data by choosing the interval midpoint? In fact, a procedure that is commonly used is to approximate y_i^* by

$$\tilde{y}_i = \frac{M_i + m_i}{2}$$

[4] We do not have a formal proof, but it should follow from consistency of $\hat{\Pi}$ and the clear fulfillment of the identification condition stated in Wooldridge [17], p. 354.

and assume that $\tilde{y}_i$ can be used as a proxy for y_i^* more or less painlessly: an additional source of error in the model (most likely heteroskedastic), that could be accommodated via robust estimation of the parameters covariance matrix. Hence, running TSLS on $\tilde{y}_i$ may look as a simple and inexpensive procedure.

Trivially, a first problem that arises with this method is that it does not provide an obvious indication on how to treat unbounded observations (that is, when $m_i = -\infty$ or $M_i = \infty$). A more serious problem, however, is that that the above procedure leads to substantial inference errors. The analytical explanation is obvious after rearranging equation (3) as

$$\tilde{y}_i = Y_i'\beta + X_i'\gamma + (\epsilon_i + \eta_i), \tag{9}$$

where η_i is defined as $\tilde{y}_i - y_i^*$. The intuition behind this reasoning is that if the interval (m_i, M_i) is "small", then σ_η^2 should be negligible compared to σ_ϵ^2. (It should be noted that, by construction, the support of η_i is a finite interval, whose length goes to 0 as $M_i - m_i \to 0$.)

However, even if the basic instrument validity condition $E(\epsilon_i|W_i) = 0$ holds, there is no reason why the midpoint rule should guarantee $E(\eta_i|W_i) = 0$. This can be proven by a simple extension to the IV case of the line of reasoning in [16]. As a consequence, the TSLS estimator converges in probability to a vector that differs from the true values of β and γ. It is worth noting that inconsistency is not a small sample issue, but a much more fundamental flaw.

Clearly, how serious the problem is depends on the relative magnitudes of σ_η^2 and σ_ϵ^2. To explore the consequences of the above, in a seemingly harmless case, we run a small Monte Carlo experiment. The Monte Carlo setup is:

$$\begin{gathered} y_i^* = \gamma_0 + Y_i\beta + X_i\gamma_1 + \epsilon_i \\ \gamma_0 = \beta = \gamma_1 = 1 \\ Y_i = 1 + X_i + Z_i + u_i \\ \begin{bmatrix} \epsilon_i \\ u_i \end{bmatrix} \sim N\left(0, \begin{bmatrix} 1 & 0.25 \\ 0.25 & 1 \end{bmatrix}\right) \end{gathered}$$

and the cutpoints are represented by the vector $[-2, 0, 1, 2, 5]$, so that, for example, if $y_i^* = 3$, then $m_i = 2$ and $M_i = 5$. The variables X_i and Z_i are independent $N(0, 1)$. A "naive" proxy for y_i^* was constructed via the midpoint rule, as

$$\tilde{y}_i = \begin{cases} -4 & \text{for } y_i^* < -2 \\ \frac{M_i + m_i}{2} & \text{for } -2 < y_i^* < 5 \\ 10 & \text{for } y_i^* > 5 \end{cases}$$

The above DGP was simulated with sample sizes of 100, 500 and 2500 observations. For each case, 4096 simulations were run.

Table 1. Monte Carlo experiment: sample size = 100

	γ_0	β	γ_1
TSLS (mean)	1.2448	1.2538	1.2480
TSLS (median)	1.2361	1.2502	1.2498
LIML (mean)	1.0009	1.0044	0.99968
LIML (median)	0.9976	1.0017	1.0025
mean estimated s.e. (TSLS)	0.2684	0.2694	0.1910
mean estimated s.e. (robust TSLS)	0.2420	0.2629	0.1872
mean estimated s.e. (LIML)	0.1736	0.1791	0.1294
Monte Carlo s.e. (TSLS)	0.2483	0.2699	0.1952
Monte Carlo s.e. (LIML)	0.1810	0.1894	0.1361
size of t-test at 5% (TSLS)	0.1133	0.1467	0.3008
size of t-test at 5% (robust TSLS)	0.1699	0.1606	0.3037
size of t-test at 5% (LIML)	0.0603	0.0635	0.0549

Table 2. Monte Carlo experiment: sample size = 500

	γ_0	β	γ_1
TSLS (mean)	1.2463	1.2505	1.2489
TSLS (median)	1.2429	1.2501	1.2500
LIML (mean)	1.0005	1.0007	0.9997
LIML (median)	1.0003	0.9995	0.9987
mean estimated s.e. (TSLS)	0.1169	0.1169	0.08273
mean estimated s.e. (robust TSLS)	0.1073	0.1170	0.08343
mean estimated s.e. (LIML)	0.0773	0.0798	0.05756
Monte Carlo s.e. (TSLS)	0.1068	0.1186	0.0836
Monte Carlo s.e. (LIML)	0.0781	0.0810	0.0585
size of t-test at 5% (TSLS)	0.5603	0.5737	0.8428
size of t-test at 5% (robust TSLS)	0.6389	0.5684	0.8369
size of t-test at 5% (LIML)	0.0527	0.0552	0.0515

Table 3. Monte Carlo experiment: sample size = 2500

	γ_0	β	γ_1
TSLS (mean)	1.2469	1.2503	1.2485
TSLS (median)	1.2478	1.2503	1.2478
LIML (mean)	1.0004	1.0005	0.9995
LIML (median)	1.0009	1.0005	0.9990
mean estimated s.e. (TSLS)	0.05193	0.05193	0.03673
mean estimated s.e. (robust TSLS)	0.04777	0.05223	0.03722
mean estimated s.e. (LIML)	0.03449	0.03564	0.02570
Monte Carlo s.e. (TSLS)	0.04744	0.05329	0.03758
Monte Carlo s.e. (LIML)	0.03458	0.03608	0.02590
size of t-test at 5% (TSLS)	0.9985	0.9971	1.0000
size of t-test at 5% (robust TSLS)	0.9990	0.9971	1.0000
size of t-test at 5% (LIML)	0.0488	0.0552	0.0515

The results are summarised in tables 1–3, which are organised as follows: the first four lines report the Monte Carlo mean and median for the two estimators. The next three lines report the mean of the estimated standard errors: for TSLS both robust and non-robust versions are reported, while the robust "sandwich" estimator[5] is used for ML. The next two lines report the ex-post dispersion of the parameters, namely the standard error of the estimates across the 4096 replications. Note that estimated and Monte Carlo standard errors should roughly match, if inference is to be at all credible. The last group of three rows shows the frequency of rejection of the hypothesis that the corresponding parameter equals its true value at 95%.

The message should be rather clear: while the LIML estimator is consistent and remarkably reliable even at a moderate sample size, the application of TSLS to the naïve "midpoint" dependent variable proxy leads to seriously inconsistent estimates and substantial inference errors.

4 An Empirical Application: the Analysis of Immigrants' Remittance Behaviour

Remittance flows are certainly one of the most interesting aspects connected to international migration, drawing in the last decades the attention of economic literature.

Macroeconomic analyses of this phenomenon are usually built on aggregate data from countries' Balance of Payments and need to take care of the fact that these measure only official flows of remittances, while the huge amounts of money transferred through unofficial channels are not taken into account. This shortcoming is less affecting survey data, where generally information on remittances are collected regardless the channel used to send them in the country of origin.

At the microeconomic level, remittance behaviour of immigrants is usually analysed as a function of migrants' characteristics and of the household's welfare in the country of origin.

Since the pioneering work of [9] on Botswana, many attemps have been made to identify the motivations to remit: altruism, inheritance, self-insurance and so forth; for an exhaustive survey of the contributions on the topic, see [13]. Remittance behaviour, anyway, could hardly be expected to depend on a single driving force, since different motivations can coexist in the same individual. Moreover, discriminative tests are empirically difficult to build for the fact that surveys seldom account for characteristics of migrants together with

[5] See for instance Davidson and MacKinnon [4], chap. 10.

information on recipient households[6], that are both essential elements to infer explanations on the motivation to remit.

The most interesting and crucial aspect in our opinion is that the empirical literature dealing with the topic usually treats migrant's income as an exogenous determinant of remittance behaviour. Yet, the need of sending money back home can affect working, consumption and possibly also investment decisions. In order to remit more an immigrant could, for example, either decide to increase the number of hours worked per week, or invest a share of his savings and make profits out of it. The amount of money to remit (if any) is therefore determined jointly in the broader context of household's strategies. Hence, in our opinion the best way to address the problem would be estimating a remittance equation that detects the main determinants of remittance behaviour addressing endogeneity and reverse causality relationships between remittances, income, consumption and saving[7].

Another central point to be noted is that, as mentioned before, data for microeconomic analyses on remittance behaviour often are taken from household surveys and it is commonly the case that in surveys' questionnaires the amount of remittances is designed as a discrete ordered variable, with different intervals mutually exclusive. If the problem is then to analyse remittance behaviour dealing with a discrete ordered dipendent variable, and addressing reverse causality between remittances, income and consumption using IV techniques, the Gretl routine just illustrated is the instrument needed to carried out our estimations.

4.1 Data and estimation issues

The dataset used in this empirical application is the Longitudinal Survey of Immigrants to Australia (LSIA), a longitudinal study of recently arrived visaed immigrants undertaken by the Research Section of the Commonwealth Department of Immigration and Multicultural and Indigenous Affairs.

We consider the first cohort of the LSIA (LSIA1), that was selected from visaed immigrants aged 15 years and over, who arrived in Australia in the two year period between September 1993 and August 1995. The sampling unit is the Primary Applicant (PA), the person upon whom the approval to immigrate was

[6] An exception is represented by the paper by [11], where migrants are considered together with their respective origin-families. Such a complete information, on the other hand, come together with a very limited number of observation, 61 pairs.

[7] [5] propose a simple theoretical model where the optimal level of remittances and savings are jointly determined. However, being mainly interested in how temporary migration affect remittance behaviour, in the empirical part they only address the possible endogeneity of the decision to come back permanently to the home country.

based. The population for the survey consisted of around 75000 PAs and was stratified by the major visa groups and by individual countries of birth.

Individuals were interviewed three times: the first time five or six months after arrival, the second time one year later and the third a further two years later[8]. Questionnaires were divided into sections and each of them is related to a different topic: migrant's family in Australia and relatives left in the country of origin, the immigration process, the initial settlement in Australia, financial assets and transfers (remittances), working status, income, consumption expenditures, education and English knowledge, health, citizenship and return visits to the former country. All these information together give an incomparable socio-economic picture of immigrants, that is essential to understand their remittance behaviour.

The sample includes 5192 individuals, but only 3752 were interviewed in all the three waves. What is relevant here is that data do not concern only a specific ethnic group, but people from more than 130 different countries[9] (both developed and developing countries). As we will highlight later, the exploitation of this cross-country dimension is an important element of the present analysis. The remittance equation that we estimate can be written as:

$$r_i = \alpha_1^* y_i + \alpha_2^* c_i + \alpha_3^* X_i + u_i$$

where r_i represents the amount of money sent home every year, y_i the yearly income of the migrants' household and c_i the total yearly consumption expenditures. r_i, y_i and c_i are all expressed in natural logarithms.

Both income y_i and consumption c_i, our endogenous variables[10], are regressed on X_i together with another set of exogenous variable, Z_i, defined as instruments:

$$y_i = \beta_1^* X_i + \beta_2^* Z_i + \epsilon_i$$
$$c_i = \gamma_1^* X_i + \gamma_2^* Z_i + \upsilon_i$$

When immigrants were asked about the amount of money sent home, they had to choose between six different intervals: 1-1000 AUS $, 1001-5000, 5001-10000, 10001-20000, 20001-50000, more than 50000 AUS $.[11] Since observa-

[8] Unfortunately, the time between interviews may vary substantially between households; this problem, together with considerable sample attrition, led us to ignore the "panel" aspect of our dataset and use all data as pooled data.

[9] As a matter of fact, the vast majority of the contributes investigate remittance behaviour of a specific nationality of migrants. Exceptions are the studies carried on using data from the German Socio-Economic Panel [10, 5, 14, 7] and the work by [3].

[10] Strictly speaking, y_i and c_i are not observed continuously either, but expressed in intervals just like the remittance variable. Not to introduce further difficulties, we take the midpoints of the intervals.

[11] There is an explicit question - section T in wave 1, section F in wave 2 and 3 - where immigrants have to answer about the amount of money sent overseas. Moreover, immigrants were

tions concentrate mainly in the first two intervals, 1-1000 AUS $ and 1001-5000 AUS $, the upper four are reduced to a single one that goes from 5001 AUS $ upwards. The final outcome is therefore a variable r_i with three possible different outcomes:

$$r_i = \begin{cases} 1 \text{ for } 1 < R_i < 1000 \\ 2 \text{ for } 1001 < R_i < 5000 \\ 3 \text{ for } R_i > 5001 \end{cases}$$

where R_i represents the real amount of money remitted. Table 4 shows the frequency distribution for the remittance variable used in the estimations.

Table 4. Remittance behaviour of immigrants in LSIA1: frequency distribution

Amount remitted	Absolute Freq.	Cumul. Freq.	%	Cumul. %
1-1000 AUS $	1584	1584	64%	64%
1001-5000 AUS $	728	2312	29.4%	93.6%
more than 5000 AUS $	164	2476	6.6%	100%

X_i includes two different sets of control variables that can influence the remittance behaviour. The first one refers to immigrants' individual characteristics: age, a dummy for the gender, a dummy for the presence of close relatives (partner, children, parents, brothers) in the country of origin, another dummy for the intention to return to the home country and the time passed since the arrival in Australia. Moreover, the level of education attained is added as a further control. Migrants may actually send money at home to repay a loan used to finance their investment in human capital. If this was the case, the higher the level of education achieved, the higher should be the amount of money sent back to the family at home. Educational attainment is divided into five levels, the first corresponding to upper tertiary education and the last to primary education.

The second set of control variables includes macroeconomic characteristics of the countries of origin. If on one hand we cannot help but recognise that the biggest shortcoming of this dataset is the complete absence of any concrete information about remittances' recipients, necessary to deal exhaustively with the motivations to remit, on the other hand the wide set of countries of origin allows

asked also about the value of assets transferred from Australia to relatives or friends overseas, in the form of personal effects, capital equipment or funds. All these transfers should be considered as remittances, especially funds, but we are not able to put them together and hence have a broader measure of remittances for the different codification of the answers fixed in the questionnaire. Hence the analysis here refers to the specific question about the money sent overseas and does not consider other transfers.

us to consider if and how remittance behaviour is influenced by macroenonomic aggregates. What we use here is first of all the level of per capita GDP, as a general measure of the level of development and wealth[12]. Secondly, the level of financial development proxied by a measure of demand, time and saving deposits in deposit money banks as a share of GDP, taken from the widely employed dataset on financial structure built by [1] for the World Bank. The rationale behind this choice is that the decision to send money back home could be influenced by the trust in the domestic financial system, especially in case of return migration when immigrants could be inclined to invest or simply save money in their home country. Finally, the distance between Australia and the country of origin is considered to proxy somehow for the costs connected to money transfers that are likely to increase the farther the homecountry is[13].

As explained before, to address problems of endogeneity both income and consumption are instrumented using a set Z_i of five instruments. The first two instrument refer to the migrant's knowledge of the English language: we use two dummy variables, one for English being the language the immigrant speaks best and another one which equals 1 if the immigrant declared a good knowledge of English; we assume that language skills should influence income prospects but not remittance behaviour. The third instrument is a dummy variable stating if the immigrant lives in a urban or a rural environment. The idea behind this choice is that the level of consumption may different between urban and rural population, but this should not affect the amount of money sent home. A dummy for child presence in migrant's household is also employed, supposing that its incidence on remittances is limited to the effects of having children on consumtpion levels. Finally, the last instrument is represented by the number of migrant's household members, expressed in natural logarithm.

4.2 Dealing with the selection problem

While estimating the remittance equation just illustrated above, we do not consider selection problems that arise when the sample is made up of people who remit and people who do not, and the variable is therefore truncated below a zero threshold. Remittances actually could be equal to zero either because immigrants are not interested in remitting to anybody, or simply because they do not earn enough to send a share of their income overseas.

[12] Data are from the World Development Indicators database.

[13] Even if we are not entering the debate on motivations to remit, geographical distance could represent in a sense also a measure of the strength of family relationship with those left behind. The source of the data employed here is CEPII (Centre d'Etudes Prospectives et d'Informations Internationales) dataset on bilateral distances.

A solution widely used in the empirical literature on the topic is the estimation approach outlined by [6], following which the decision to remit is modeled as a two-stage sequential process. In our case, an extension of sample selection models à la Heckman that involves interval estimation as a second step instead of OLS does not exist in the literature, to the best of our knowledge, and is certainly not straightforward to conceive and implement. All the more so, when instrumental variable interval estimations are needed.

We have made a rough attempt to redesign the remittance variable adding one initial class that includes people who send zero AUS$ in the country of origin. These observations are singled out looking if at the yes/no question whether they remit, immigrants gave a negative answers. If this is the case, the observation is considered as zero.

The main idea behind this strategy is that, if the population were homogeneous, the problem coul be dealt with simply by considering the non-senders as having $-\infty < y_i^* \leq 0$ (in the language of equation (3)). In this scenario, the only possible reason for not sending money abroad is the budget contraint. If, on the other hand, there are people who would not send remittances whatever their income, then the two-stage decision process should be modelled separately.

Running again estimations with the zero-augmented dependent variable we get results that are quite different from the original ones, both in terms of significance and in terms of magnitude. This has to be read as a clear sign that sample selection is a problem we absolutely have to deal with, but at the same time it also shows that a suitable tool is needed. Considering a class of zeros *de facto* does not address correctly the selection mechanism, because we use a common model for the two different group of individuals and the estimation of a remittance equation actually does not make much sense for an immigrant who is not interested in sending money back home.

We prefer therefore not to introduce a strong source of heterogeneity by joining two samples (remitters and non remitters) that are most likely to be structurally different; we are aware that our results should be considered as conditional on the fact that the individual is a remitter.

The next step in our research then will be to include in the model a selection equation to control properly for the selection mechanism, but meanwhile the main focus of the work is on endogeneity treatment and the adoption of IV technique in interval estimations.

4.3 Results

Results are reported in Table 5. The first three columns show results obtained with simple interval estimations, while from column 4 onward we introduce IV techniques.

Table 5. Estimates for the Australian remittances data

	Non-IV			IV		
	[1]	[2]	[3]	[4]	[5]	[6]
const	-1.50	2.04	1.53	**13.03**	**17.35**	**17.48**
male	**0.26**	**0.31**	**0.31**	*0.24*	*0.29*	*0.32*
age	0.00	0.00	0.00	0.00	0.00	0.00
time in AUS	**0.50**	**0.49**	**0.48**	**0.51**	**0.51**	**0.51**
back home	*0.37*	*0.40*	*0.42*	*0.40*	*0.43*	*0.44*
relatives overseas	0.11	0.04	-0.09	0.03	-0.04	-0.08
qualifications_2	0.24	0.17	*0.33*	0.21	0.13	0.18
qualifications_3	-0.13	-0.16	-0.03	-0.19	-0.24	-0.16
qualifications_4	*-0.38*	*-0.43*	-0.24	-0.14	-0.24	-0.14
qualifications_5	**-0.60**	**-0.66**	**-0.54**	**-0.62**	**-0.71**	**-0.61**
per capita GDP	*0.12*	**0.19**	0.08	**0.22**	**0.29**	0.16
deposit			*0.14*			*0.16*
distance		**-0.41**	*-0.29*		**-0.43**	*-0.27*
income	**0.24**	*0.22*	*0.20*	**1.13**	**1.00**	**1.15**
consumption	**0.36**	**0.36**	**0.43**	**-2.16**	**-2.12**	**-2.31**
N	1136	1135	983	1132	1131	979
σ	1.20	1.19	1.19	1.46	1.43	1.48
Wald test				15.22	17.00	15.06
Wald test p-value				0.00	0.00	0.00
Over-id. test				9.75	7.31	3.98
Over-id. test p-value				0.02	0.06	0.26

Note: coefficients in **boldface** are significant at 1%; coefficients in *italics* are significant at 5%; coefficients in normal fonts are significant at 10%; coefficients in small fonts are not significant.

When not instrumented, both income and consumption are statistically significant with a positive sign. The result is expected for income, and in line with the previous findings of the empirical literature[14], but quite puzzling for consumption. If we consider remittances as a sort of savings, the natural prediction would be that they diminish as the level of consumption expenditure of the immigrants' household increases. This clearly shows how results are biased when we do not take reverse causality into account.

Moving to IV interval estimations (column 4-6), income and consumption are statistically significant at 1%, the former with a positive sign and the latter negatively. Elasticity of remittances to income is around 1-1.15, while elasticity to consumption is slightly bigger than 2. Remittances seem therefore much more responsive to change in consumption expenditure compared to change in the level of income.

Among individual characteristics, the age of immigrants seems not to influence their remittance behaviour, while gender differences result statistically significant: other things being equal, male migrants remit on average 30% more than female. The desire to return living in the country of origin predictably affects the amount remitted in a significant way, with potential returnees remitting around 40% more. Time elapsed from the arrival in Australia has also a positive and significant effect.

The presence of a close relative still living overseas does not play such a significant role in determining remittances. This somehow surprisingly result could be due to the fact that in the sample considered here almost everybody who sends money overseas has at least one close relative (spouse, children, parents, brothers/sisters) still living in the country of origin.

As far as the immigrants' education is concerned, just one out of four dummies is significant across all the specifications, with a negative sign, and is the one associated to the lowest level of education (primary school). What emerges is hence that, even after controlling for the level of income, more educated migrants are likely to remit higher amounts than the less educated.

The Wald test rejects firmly the exogeneity hypothesis for income and consumption. Endogeneity effects are therefore highly significant, so specification 1-3, which do not take this into account, must be regarded as incorrect. If covariances between residuals from the first steps and residuals from the remittance equation are considered, it is clear that the result of the Wald test is driven mainly by consumption that is strongly endogenous, while income is understandably less affected from problems of reverse causality. Moreover, the result from the LR test of over-identifying restrictions confirms the validity of the set

[14] Among others, see [2] and [3].

of instruments we have chosen to address endogeneity of income and consumption, especially in the complete specification.

As explained before, the cross-country dimension is taken into account adding to individual characteristics macroeconomic variables concerning the home countries. Surprisingly, per capita GDP of immigrants' country of origin turns out to be significant with a positive sign. Immigrants coming from richer countries seem to remit more. This result is confirmed when we consider also the distance between Australia and immigrants' home country (column 5).

The most interesting aspect is that per capita GDP loses all its explanatory power when the level of financial development is added as a further explanatory variable (column 6), while financial development is significant at 5% and plays an important role in immigrants' household decision. Per capita GDP hence seems to act, when significant, as a sort of proxy for the level of financial development of immigrants' country of origin, but this is indeed the macroeconomic feature that matters when immigrants consider how much money to remit.

5 Conclusions

We argue that estimation of models in which the dependent variable is observed by intervals and explanatory variables may be endogenous ought to be conducted via maximum likelihood, all the alternative possibilities being inefficient at best and plain wrong at worst.

An example with Australian remittances data shows that our procedure is effective. Endogeneity of income and consumption in the context of immigrants' remittance behaviour does matter. Consumption is strongly endogenous, while income is less affected from problems of reverse causality; anyway, endogeneity effects are altogether highly significant and need to be addressed in empirical models. Failing to account for them will lead to incorrect estimates.

Bibliography

[1] BECK, T., A. DEMIRGUÇ-KÜNT, AND R. LEVINE (2000): "A New Database on Financial Development and Structure," *World Bank Economic Review*, 14, 597–605.

[2] BROWN, R. P. (1997): "Estimating Remittance Functions for Pacific Island Migrants," *World Development*, 25(4), 613–626.

[3] CLARK, K., AND S. DRINKWATER (2007): "An Investigation of Household Remittance Behaviour; Evidence from the United Kingdom," *The Manchester School*, 75(6), 717–741.

[4] DAVIDSON, R., AND J. G. MACKINNON (1999): *Econometric Theory and Methods*. Oxford University Press, Oxford.

[5] DUSTMANN, C., AND J. MESTRES (2008): "Remittances and Temporary Migration," mimeo.

[6] HECKMAN, J. J. (1979): "Sample Selection Bias As a Specification Error," *Econometrica*, 47(1), 153–161.

[7] HOLST, E., A. SCHAEFER, AND M. SCHROOTEN (2008): "Gender, Migration, Remittances : Evidence from Germany," SOEPpapers 111, DIW Berlin, The German Socio-Economic Panel (SOEP).

[8] KAWAKATSU, H., AND A. G. LARGEY (2009): "EM Algorithms for Ordered Probit Models with Endogenous Regressors," *Econometrics Journal*, forthcoming.

[9] LUCAS, R. E., AND O. STARK (1985): "Motivations to remit: evidence from Botswana," *Journal of Political Economy*, 93, 901–918.

[10] MERKLE, L., AND K. F. ZIMMERMANN (1992): "Savings, Remittances and Return Migration," *Economics Letters*, 38, 77–81.

[11] OSILI, U. O. (2007): "Remittances and Savings from International Migration: Theory and Evidence using a Matched Sample," *Journal of Development Economics*, 83, 446–465.

[12] PAGAN, A. (1986): "Two Stage and Related Estimators and Their Applications," *Review of Economic Studies*, 53(4), 517–538.

[13] RAPOPORT, H., AND F. DOCQUIER (2005): "The Economics of Migrants' Remittances," IZA Discussion Papers 1531, Institute for the Study of Labor (IZA).

[14] SINNING, M. (2007): "Determinants of Savings and Remittances: Empirical Evidence from Immigrants to Germany," IZA Discussion Papers 2966, Institute for the Study of Labor (IZA).

[15] SOLIÑO, M., M. X. VÁZQUEZ, AND A. PRADA (2009): "Social demand for electricity from forest biomass in Spain: Does payment periodicity affect the willingness to pay?," *Energy Policy*, 37(2), 531–540.

[16] STEWART, M. B. (1983): "On least squares estimation when the dependent variable is grouped," *Review of Economic Studies*, 50(3), 737–753.

[17] WOOLDRIDGE, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, Mass.

Applied Econometrics

A Model for Pricing the Italian Contemporary Art Paintings at Auction

Nicoletta Marinelli[1] and Giulio Palomba[2]

[1] Dipartimento di Economia, Università Politecnica delle Marche, Italy. *n.marinelli@univpm.it*
[2] Dipartimento di Economia, Università Politecnica delle Marche, Italy. *g.palomba@univpm.it*

Abstract. This paper aims to model the auction prices of Italian contemporary art paintings. The contribution to the existing literature is twofold concerning both the methodological and the conceptual aspects. From the former point of view, we use the two-stages Heckit model which allows us to take into account the sample selection bias deriving from the "buying" risk, that affects transactions at auction. From the latter point of view, we have found that some sale characteristics such as auction house prestige and year of sale, are more important than the physical aspects of the paintings. Moreover, some artistic characteristics, the artist's name and their living status are also relevant.
The whole analysis is carried out after creating a new dataset of 2817 transactions which took place at the most important auction houses between 1990 and 2006.

1 Introduction

The prices of paintings depend upon a set of variables, concerning the characteristics of the paintings themselves, but also other aspects more difficult to be measured, such as the artist's popularity or the auction house's prestige. Several questions about the level of art prices are still open and literature has not clearly defined what are the main drivers of their dynamics and what are the conditions for a more liquid and riskless investment in artworks.

From the theoretical point of view, there are two main theories regarding the price determination: on one hand, [3] claims that there may exist no equilibrium level for art prices, so they can float more or less aimlessly with unpredictable oscillations emphasized by the activities of investors/speculators; on the other hand, [17] assume that a "natural price" does not exist for paintings, neverthless market forces related to demand and supply determine prices for artworks, as for any other economic good.

From the empirical point of view, the pricing of paintings is generally discussed within the framework of market price indexes, with the aim of evaluating the rate of return of an investment upon such assets. In this context, the hedonic regression (from [2] onwards) seems to be a good methodology to select the variables which can be useful to model the evolution of artwork prices.

The key-objective of this paper is to carry out an empirical analysis about the price determinants of Italian contemporary art paintings at auction. The analysis is two-fold because it allows us to jointly model how some explanatory variables contribute to the probability of having an unsold item and to the price levels of sold works.

In doing so, a preliminary sample selection is obviously required. We consider a sample of 2817 painting transactions from the 21 Italian contemporary artists who showed the biggest turnover at auction during the period 1998-2002, according to [35]. Starting from the available information about this sample of transactions, we made a new dataset in which all the variables are grouped into four categories, being the usual painting-specific attributes: they are the physical qualities of the work, the characteristics of the artist, the artistic and the sale characteristics of the paintings.

The reminder of this paper is organized as follows: in section 2 we introduce the problems related to the sample selection and the choice of the relevant variables. The whole empirical analysis is carried out in section 3 and section 4 concludes. Finally, the Appendix includes the complete list of all available variables.

2 The sample selection

The analysis of the price dynamics of paintings sold at auction has to be based upon the choice of an appropriate sample. In this article, all the available information is taken from "Artindex Plus", a detailed database which contains the catalogue's information about several artworks[3]: more precisely, it provides the picture of the painting plus different pieces of information about the artist and the artwork itself (see section 2.1 for details).

Our sample choice substantially depends upon the reaching of a sort of homogeneity between variables: given that the market of paintings is composed of unique goods, we focus the attention upon Italian contemporary art because we need to deal with goods as comparable as possible[4].

Since Italian contemporary art itself is not completely homogeneous[5], we limited our analysis to the 21 Italian contemporary artists who showed the biggest turnover at the most important international auctions during the period 1998-

[3] Artindex Plus is provided by Gabrius S.p.A. operating in Milan and belonging to the Munus Culture Holding (AMB network); for more details see *http://www.munusartinvest.com.*

[4] The market of paintings is usually divided in four branches which have their own dynamics and characteristics: Old Master, XIX Century, Modern Art and Contemporary Art.

[5] In practice, there are differences among "emerging" and "historical" contemporary art painters.

2002, according to [35][6]. The reason for this selection is that the paintings are considered as investment goods for which the main characteristics depend upon the market dynamics; the aestethic component is not supposed to be relevant here. The homogeneity in our sample is also preserved by the exclusion of prints and drawnings because these items have their own specific price dynamics, as claimed by [26], and are often traded in separate sessions at auction.

Finally, we restrict the period of observation to the years which go from 1990 to 2006, since the Artindex Plus data regarding auction sales before 1990 are very poor and incomplete. Following this sample selection, we work with a dataset of 2817 painting transactions placed at the most important auction houses.

A problem encountered in studying art prices stems from the fact that the auction data samples could suffer from some problems of selection bias, as already underlined by [38]. It is well known that the art market is divided into "primary", "secondary" and "auction" market: in the former the artist personally sells her works to buyers, while in the second the galleries and the art dealers trade paintings with private or institutional collectors. Auction represents the remaining solution, therefore it can not take into account all types of paintings. Neverthless, in this case public information exists and this overcomes most of the typical problems due to the incomplete and asymmetric information availability of the art market. Moreover, we suppose that auction prices affect the art market because collectors and professional art dealers take these price as guidelines, following the approach of [16]. Finally, we also consider auction prices as adequate approximations of true equilibrium prices, as pointed out by [6].

With this sample selection, we try to give an empirical contribution for a sector that literature has often neglected[7].

2.1 The data

For each item Artindex Plus provides the following information: a picture of the painting, personal details about the artist, physical characteristics of the painting (date of execution, width and height, support, medium), artistic characteristics

[6] [35] define the "turnover" as the number of sold works multiplied by their mean price.

Moreover, they conventionally define as the Italian contemporary artists those Italian painters who carried out their activity after 60's. This selection criterium is not strictly applied, since some Italian painters, still working after 1960's, but historically placed with the best artists of Futurism or other artistic currents preceding the 1960's, are not included in their sample (for example, Carlo Carrà). So, in the analysis of [35], the Italian contemporary art conventionally starts with the contributions of Fontana (1899-1968), Burri (1915-1995), Marini (1901-1980) and Manzoni (1933-1963).

[7] For previous contribution see for example [5], [1], [30], [34] or [25]. Only [6] uses data about the Italian market of Modern and Contemporary oil paintings.

of the painting (list of previous owners, signature, date, title, expertise, literature citations, list of exibitions), sale characteristics of the painting (lot number, auction house, city, month and year of transaction), economic characteristics of the paintings (hammer prices, hammer prices plus transaction fees, pre-sale evaluation by experts who provide the estimation of a range of prices).

Tables in the Appendix report the descriptions of the variables used in our work.

2.2 Dependent variables

Given that we aim to model the auction price levels taking into account the problem of unsold paintings, our dependent variable is given by the auction price of each painting. In our dataset we have both the hammer price and the total purchase price: the latter differs from the former because it includes the auction house's transaction fees. All the prices related to unsold paintings at auction are not observable, hence they are set as zero.

Both types of prices are all converted to US Dollars to make them comparable, obtaining series p_i and P_i respectively. Finally, we consider their logarithmic transformation, indicated with y_i and Y_i.

2.3 Explanatory variables

The main evidence related to the variables identification concerns the qualitative nature of most of the available data; for this reason several variables are dummies. The explanatory variables for the price of Italian contemporary art paintings are organized into four categories; the list of potential price determinants and their codes are reported in the Appendix.

A. Characteristics of the artist: personal characteristics of the artist who painted the work.
- *1) Name of the artist*: 21 different dummy variables, one for each artist in the sample.
- *2) Living status*: dummy variable[8] (1 if the painter is deceased at the time of the sale and 0 otherwise).
- *3) Year of birth*.

B. Physical characteristics: related to the execution of the artwork.
- *4) Medium*: this variable allows us to control the assumption of a superior market value as a consequence of the media durability and particulars[9].

[8] All other things being equal, the price of artworks are often supposed to increase once an artist has died, as pointed out by [24].

[9] Generally, oil paintings are supposed to be more expensive than other media. See, among others, [11], [34], [25], [24].

5) *Support*: 10 different types of support upon which the artwork is painted are available. The related dummy variables have the value of one when the specified support is used, alone or jointly with another, and zero otherwise.
6) *Size*: the surface (expressed in m^2) and the squared surface as in [34], [24] and [40]. In particular, [11] describes the price of painting as a concave function of dimensions.

C. Artistic characteristics: these variables are supposed to be as proxies of the prestige and the popularity of the artwork in the art world. They are all dummy variables taking into account for the following characteristics:

7) *Authentication by the artist*
8) *Publication in catalogues or monographies*
9) *Date*
10) *Recognition by experts*
11) *Literature*: citations in the artistic literature (see [14])
12) *Signature*
13) *Title*
14) *Exhibitions*
15) *Number of previous owners*: according to auction houses, the price reached for a painting is influenced by its provenance. The number of previous owners can be useful in order to test whether a painting rarely traded in the auction market reaches a greater price than a painting that has often been put on sale (see [15]). Obviously, this is not a dummy variable[10].

D. Sale characteristics: with this set of explanatory variables we test the hypothesis that sale conditions have an effect upon the marketability and upon the final price reached by the painting at auction.

16) *Auction house*: [32], [13], [34], [25], [24], among others, show that Christie's and Sotheby's systematically obtain higher hammer prices; this evidence is generally attributed to the leading role played by both institutions in this business.
17) *Marketplace*: dummy variables for the 18 different marketplaces in database.
18) *Sale date*: dummy variable for each year (from 1990 to 2006) and for each month of sale.
19) *Pre-sale estimates*: before an auction sale takes place, experts usually provide an estimate of the potential market value of the painting. Pre-sale estimates are usually provided as a range.

[10] The dataset does not allow us to classify all previous owners according to their institutional nature (for example, museum, gallery or private collector), because it provides only the names of previous owners.

3 The model

The aim of our proposed methodology is to model the auction prices of the Italian Contemporary Art paintings. Examining the determinants of auction prices from a speculative perspective, we have to consider the possibility of unobserved final prices; in other words, as well as the price reached by sold works, we have to take into account the "buying risk" affecting each transaction. Since in our sample various artworks go unsold, the analysis must be divided into two stages: in the first stage, a distinction between sold and unsold paintings is made, while in the second stage, prices of sold paintings are modelled.

3.1 The Heckit model

From the statistical point of view the possibility of unsold items at auction imply a problem of selection bias which can arise from censoring data. In particular, the properties of painting prices can vary taking unsold works into account, thus data can suffer from nonrandomness.

To address this problem the Heckit model [23] is used; this model allows us to carry out the analysis when the dependent variable is continous but censored for values under a defined threshold. This methodology was introduced to correct the selection bias occurred for nonrandomly selected samples and provides consistent estimates which eliminate the specification error for the case of censored data. Recently, [40] used this methodology upon a sample of Picasso prints censored for repeat-sales, as well as [8] upon a sample of Symbolist paintings.

Analitically, the Heckit model consists of

$$\begin{cases} s_i^* = z_i'\gamma + u_i \quad i = 1, 2, ..., N \\ w_i = x_i'\beta + \varepsilon_i \;\Leftrightarrow\; s_i^* > 0, \end{cases} \tag{1}$$

where N is the sample size. The first equation is the "selection equation", where s_i^* is a latent variable which is positive if the auction price is greater than the reservation price. Moreover, the $1 \times K$ vector z_i' contains the individual characteristics that determine if the painting is sold or not, γ is a K-dimensional vector of unknown parameters and u_i is a random disturbance. The latent variable s_i^* is not observed, therefore we define a dichotomic variable s_i as

$$s_i = \begin{cases} 1 \;\text{ if } s_i^* > 0 \\ 0 \;\text{ otherwise.} \end{cases} \tag{2}$$

In practice, for sold paintings $s_i = 1$, while it is zero otherwise.

The second equation of the system (1) is the linear model of interest in which w_i is the dependent variable; x_i is the $1 \times M$ vector of exogenous variables, β

is a M-dimensional vector of unknown parameters and ε_i is a random error term. The explanatory variables in x_i could be also included in z_i and viceversa. Moreover, we assume that the random disturbances are jointly distributed as

$$\begin{bmatrix} u_i \\ \varepsilon_i \end{bmatrix} \sim \text{i.i.d.} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \sigma_{u\varepsilon} \\ \sigma_{u\varepsilon} & \sigma_\varepsilon^2 \end{bmatrix} \right). \tag{3}$$

In our model the selection bias arises because the price w_i is observed only when the i-th painting is sold (therefore $s_i = 1$) and when $\sigma_{u\varepsilon}$ is different from zero; in such a situation, [23] shows that OLS estimation yields biased and inconsistent estimates of β.

Generally, the estimator for the Heckit model is the Maximum Likelihood (ML) under the assumption of joint normal distribution in equation (3); this method guarantees consistent and asymptotically normal and efficient estimates (see for example [22]). Unfortunately, in our analysis ML estimation of the model (1) does not achieve convergence, hence we use the Heckman's (1979) two-step procedure which yields a less efficient estimator.

The whole procedure can be briefly outlined as follows: given that $\phi(\cdot)$ and $\Phi(\cdot)$ respectively are the density and the cumulative density functions of the standardised Gaussian distribution, the first step consists of the ML estimation of the probit model $\Pr(s_i = 1) = \Phi(z_i'\gamma)$. This equation predicts whether an item goes sold/unsold and it is useful to obtain the inverse of the Mills Ratio given by $\lambda_i = \phi(z_i'\gamma)/\Phi(z_i'\gamma)$, which will be used as an additional regressor during the second step to correct the potential sample selection bias.

Once λ_i is inserted in x_i vector, its coefficient is $\beta_\lambda = \sigma_{u\varepsilon}\sigma_\varepsilon^2$ and the second equation in (1) can be estimated via the OLS method. The covariance between u_i and ε_i can also be estimated and the standard t-statistic on β_λ is used to test if any problem of selection bias occurrs in our analysis.

Moreover, as shown by [31], the assumption of normality of the probit residuals u_i is required to have consistency and plays a key role because it represents the sufficient condition to define λ_i as in equation given above. Following [12] we carried out the following conditional moment (CM) test based on the OPG Regression[11]

$$\iota = \hat{\gamma} Z + \hat{b} G + \text{residuals}, \tag{4}$$

where ι is a vector of ones, Z is the matrix whose each row is z_i' and $\hat{\gamma}$, $\hat{b}$ are ML estimates from the probit $\Pr(s_i = 1) = \Phi(z_i'\gamma + G_i'b)$. To take into account for asymmetry and kurtosis, the i-th row of the matrix G is

$$G_i' = \left[[(z_i'\hat{\gamma})^2 + 2]\hat{u}_i \;\; z_i'\hat{\gamma}[(z_i'\hat{\gamma})^2 + 3]\hat{u}_i \right], \tag{5}$$

[11] Outer Product Gradients Regression; see for example [12] for details.

where $\hat{u}_i$ are the model generalised residuals [see for example 31]. It can be shown that, for each observation, G_i' contains the sample counterparts of the orthogonality conditions about the conditional moments $E(u_i^k | u_i < -z_i'\gamma)$, when k is 3 and 4 respectively [see 36].

The basic idea is that, if G_i' is not statistically relevant in the selection equation, the probit model is correctly specified. Hence, the null hypothesis of the CM test is $H_0 : b = 0$ and the test statistic is given by N times the R^2 of the regression (4). Given that G has two columns, the asymptotic distribution is the standard χ_2^2.

3.2 Empirical results

The starting point of our analysis consists of the Heckit estimation where the second equation in (1) can be thought as a sort of an hedonic regression in which the selection bias has been taken into account. All results for $w_i = y_i, Y_i$ are provided in Tables 1 and 2, while Table 3 contains some regression statistics; some explanatory variables among those presented in section 2.3 are dropped to avoid collinearity and, after some preliminary estimates, other variables are excluded to reach the possible maximum reduction of parameters, without any loss of relevant information.

The first emerging aspect is that the estimates of the auction prices of the Italian Contemporary Art paintings are quite similar for the logarithms of the hammer price (y_i) and of the total purchase price (Y_i): the presence of transaction fees does not seem to have any relevant impact upon our analysis, also considering that there are 4 missing values for P_i in our original sample (see the total observations in Table 3). The sample size reduction is due to three missing values in $surface$ and $squares$.

The null hypothesis of the CM test is strongly accepted in both cases and this supports the consistency of our estimates in which λ_i is not statistically different from the inverse of the Mills Ratio.

The t-statistic evaluated for λ_i indicates that some correction for the sample selection bias is needed and, for this reason, the Heckit model is superior to OLS.

The negative estimated value of the coefficient related to λ_i depends upon $\hat{\sigma}_{u\varepsilon} < 0$: this suggests that paintings that go sold are more likely to be those with a lower price, since cheaper paintings are likely to be bought by a wider group of potential buyers.

Moreover, the [4] normality test highlights that the model disturbances are not jointly normally distributed and this is probably the reason why the ML estimation process does not converge.

The contributions given by the explanatory variables in the two steps of the estimation are discussed below.

First step Only the dummies related to painters Boetti, Campigli, Fontana and Magnelli positively contribute to the outcome of artwork transactions. This suggests that the paintings made by this group of artists are, on average, less likely to go unsold at auction, showing a strong tendence to be easily traded. If the artist is dead at the moment of sale the painting has a higher probability to go unsold, as highlighted by the negative and significant coefficient related to $dead$. The variable $birth$ has been dropped according to the results of preliminary analysis in which it was found to be not statistically relevant in both steps of estimation.

Media and support do not play any relevant role upon the probabiliy that paintings go unsold; only items painted with $enamel$ are less likely to be sold. Even if in our sample most of the paintings are made on canvas and paper (see Table 5), they do not affect the estimation.

All the variables used to capture the prestige and the popularity of the paintings do not seem to be relevant at this stage of the estimation, with the only exception being $literature$ which has a very feeble effect (the p-value is about 0.11).

The outcome of the sale, in terms of sold/unsold work, is highly determined by the auction house where the sale is arranged. For the need of parsimony, we consider only Christie's, Sotheby's and Finarte where more than 90% of transactions are placed. All their coefficients are positive and highly significant. The findings about Christie's and Sotheby's are coherent with those of [13] who argued that some auction houses are able to systematically influence the successful outcome of the sale since they often attract more high valued artistic works[12]. The result of Finarte could be intepreted as a consequence of the "home bias effect", that is a general preference of buyers for domestic art production, as pointed out by [7].

It has also been previously proved that the other auction houses, the city and month of sale do not seem to have an additional effect on the probability of going unsold.

Some years affect the outcome of the sale more than others: in particular 1993, 1997 and 2002 show negative and statistically significance relationships, while 1992 and 2004 instead have a positive and significant parameter.

[12] *...the quality of a painting , not captured by our characteristics, is partly picked up by the saleroom coefficients: a "good" Picasso would go to Christie's or Sotheby's New York, a less good one would be sold at Drouot's [...] it is impossibile to disentangle the two effects.*

Second step It is straightforward evident from Table 2 that almost all variables play a key role in determining auction prices and the impact given by the majority of painters seems to be decisive. The number of the exceptions is very small and the statistical significance attributed to Pomodoro is scarse probably because only two works belong to our sample (see Table 4).

The estimation highlights that Campigli and Fontana, who have a positive and significant impact upon the selection equation, also show an analogous effect upon the second step; on the contrary, the coefficient related to Boetti has the opposite sign of that in the selection equation. The paintings made by Burri, Cattelan, Manzoni and Marini also seem to reach market values higher, on average, than other artists, while the negative parameters related to different painters suggest that their works generally achieve lower prices.

The variable $dead$ do not have any effect, while the variable $birth$ is dropped because of its statistical irrelevance. From our model one can argue that the death of the artist before the moment of sale only increases the probability that paintings go unsold, but does not affect auction prices. This result is in contrast with both contributions of [1] and [39]: the former paper showed an increase by 154% of the auction prices of American art when the artist was still alive, while the latter work found that paintings made by deceased artists are associated with a price increase of 100.58%.

Our estimation suggests that painting media do not have a relevant effect upon the total purchase price of a painting; the only exceptions are oil and the residual variable $other$ for which the coefficients produce an increasing effect upon artwork prices. It is difficult to compare these findings with previous analyses especially because these contributions are sometimes limited to historical periods when only few media were known [see for instance 14] or restricted to single medium samples [6].

The contributions of the supports are heterogeneous because $canvas$ seems to have a significant and positive influence upon painting prices, while $paper$ has the opposite effect.

The coefficient signs of the variables regarding the size of paintings are those expected and coherent with the findings of [11]: in particular, the artwork prices can be described as a concave function in which the surface and the squared surface have a positive and negative relationship respectively. This suggests that, if the size is augmented, the Italian Contemporary Art prices tend to increase at first, but then decrease when the painting becomes too large and difficult to hang.

Among the artistic characteristics of the paintings, the publication in catalogues, the number of exhibitions, the literature and the number of previous owners have a positive effect, while the variable $expertise$ surprisingly shows

negative contributions, contrary to our expectation. Variables $authentic$ and $signature$ do not have any effect upon the estimation, maybe because the prestige of some auction houses serves as a guarantee of authenticity.

The sale year substantially affects the final purchase price of Italian contemporary art paintings: each year from 1991 to 2004 shows statistically significant and negative coefficients, while years 2005 and 2006 do not seem to be relevant. From the economic perspective the series of these coefficients can be used to built the yearly price index I_t, with all other characteristics being equal. This index shows the contribution to auction prices dynamics given by years of sale and its equation is $I_t = 100 \cdot \exp\{\hat{\beta}_t\}$, where $t = 1991, 1992, \ldots, 2006$. Just the hammer price index is plotted in Figure 1 since the curve related to the total purchase price (Y_i) is very similar. The base year is 1990 in which $I_t = 100$. For both series this index substantially shows an increase from 1994, while only in 2006 it has a value greater than those of the base year. This is consistent with the evidence of the art market downturn experienced in the early ninetees [see, among others, 29] and the upturn of the market in recent years.

Finally, even if Table 1 highlights that principal auction houses strongly determine the outcome of sale, their contribution to price levels is not relevant and, for this reason, they have been dropped from the second the step of the estimation.

4 Concluding remarks

This paper aims to model the prices of paintings given a set of explanatory variables regarding different characteristics. The whole analysis is carried out after creating a sample of 2817 transactions of paintings made by 21 Italian contemporary painters and sold at auction during the period 1990-2006. To take the problem of sample selection bias arising from the inclusion of unsold paintings into account, the Heckit model [23] is used to obtain consistent estimates.

Our estimation highlights that some mechanism of selection bias occurrs hence this methodology is superior to OLS. The main finding is that auction prices for the Italian Contemporary Art market depend upon several variables such as auction house prestige, year of sale, artist's popularity and different artistic characteristics of paintings (publication in catalogues, number of exhibitions, citations in the artistic literature, number of previous owners). This finding is consistent with the main existing literature.

Contrary to previous studies [see for example, 11 or 39], we found that traditional media, supports and conventional proxies of artistic qualities are less able to explain the marketability of paintings, while they have a strong effect on price levels. Other variables playing a leading role upon the outcome of sale are

those related to sale characteristics (for example, auction house prestige) and to the years in which the transactions take place; the years of sale also affect the auction price determination.

A price index that fits the cyclical nature of the Italian Contemporary Art market has been derived from the coefficients related to the years of sale: after an initial decline it tends to increase from 1994 and finally have a strong rise after 2003. This evidence reflects the downturn of the art market in the early ninetees and it is coherent with previous literature. This is also consistent with the upturn in contemporary painting prices experienced in recent years. Some suggested reasons for this cycle could be macroeconomic factors such as the dependence of the art market upon per capita income [17], financial courses such as the correlation between art market cycles and bullish/bearish financial markets [9] or simply art fads such as collectors' changing attitudes towards contemporary art [5].

Bibliography

[1] Agnello, R. and Pierce, R.: Financial returns, price determinants and genre effects in American art investment. Journal of Cultural Economics **20** (1996) 359–383

[2] Anderson, R.C.: Paintings as an investment. Economic Inquiry **12** (1974) 13–26

[3] , Baumol, W.J.: Unnatural value: or art investment as floating crap game. American Economic Review **76** (1986) 10–14

[4] Box, G.E.P. and Pierce, D.A.: Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. Journal of the American Statistical Association **65** (1970) 1509–1526

[5] Buelens, N. and Ginsburg, V.: Revisiting Baumol's "art as a floating crap game". European Economic Review **37** (1993) 1351–1371

[6] Candela, G. and Scorcu, A.E.: A price index for art market auctions. An application to the Italian market for modern and contemporary oil paintings. Journal of Cultural Economics **21** (1997) 175–196

[7] Candela, G. and Scorcu, A.E.: *Economia delle arti*. Zanichelli, Bologna (2004)

[8] Collins, A., Scorcu, A.E. and Zanola, R.: Sample selection bias and time instability of hedonic Art Price Indexes. Working Paper DSE **610** (2007)

[9] Chanel, O.: Is art market behaviour predictable?. European Economic Review **39** (1995) 519–527

[10] Chanel, O., Gérard-Varet, L.-A. and Ginsburgh, V.: The relevance of hedonic price indices. The case of paintings. Journal of Cultural Economics **20** (1996) 1–24

[11] Czujack, C.: Picasso paintings at auction, 1963-1994. Journal of Cultural Economics **21** (1997) 229–247

[12] Davidson, R. and MacKinnon, J.D.: *Estimation and inference in econometrics*. Oxford University Press, New York (1993)

[13] De la Barre, M., Docclo, S. and Ginsburgh, V.: Returns of Impressionist, Modern and Contemporary European Painters, 1962-1991. Annales d'Economie et Statistique **35** (1994) 143–181

[14] Figini, P. and Onofri, L.: *Old master paintings: price formation and public policy implication*. in Marciano, A. & Josselin, J. M. (eds.), *The law and economics of the Welfare State: a political economics approach*, Edwar Elgar Publisher (2005)

[15] Fiz, A.: *Investire in arte contemporanea*. Franco Angeli, Milano (1995)

[16] Frey, B.S. and Pommerehne, W.W.: Art investment: an empirical inquiry. Southern Economic Journal **56** (1989) 396–409

[17] Frey, B.S. and Pommerehne, W. W.: *Muse e Mercati. Indagine sull'economia dell'arte*. Il Mulino, Bologna (1991)

[18] Ginsburgh, V. and Jeanfils, P.: Long-term comovements in international markets for paintings. European Economic Review **39** (1995) 538–548

[19] Goetzmann, W.N.: Accounting for taste: art and financial markets over three centuries. The American Economic Review **83** (1993) 1370–1376

[20] Goetzmann, W. and Spiegel, M.: Private value components and the winner's curse in Art Index. European Economic Review **39** (1995) 549–555

[21] Grampp, W.D.: *Pricing the priceless. Art, Artists and Economics*. Basic Books, Inc., New York (1989)

[22] Greene, W.H.: Sample selection bias as a specification error: comment. Econometrica **49** (1981) 795–798

[23] Heckman, J.J.: Sample selection bias as a specification error. Econometrica **47** (1979) 153–161

[24] Higgs, H. and Worthington, A.: Financial return and price determinants in the australian art market, 1973-2003. The Economic Record **81** (2005) 113–123

[25] Hodgson, D. and Vorkink, K.P.: Asset pricing theory and the valuation of canadian paintings. Canadian Journal of Economics **37** (2004) 629–655

[26] Holub, H. W., Hutter, M. and Tappeiner, G.: Light and shadow in art price computation. Journal of Cultural Economics **17** (1993) 49–69

[27] Locatelli-Biey, M. and Zanola, R.: The market for Picasso prints: an hybrid approach. Journal of Cultural Economics **29** (2005) 127-136

[28] Locatelli-Biey, M. and Zanola, R.: Investment in paintings: a short run price index. Journal of Cultural Economics **23** (1999) 211–222

[29] Mei, J. and Moses, M.: Art as an investment and the underperformance of masterpieces. American Economic Review **92** (2002) 1656–1668

[30] Mok, H.M., Ko, V.W., Woo, S.S. and Kwok, K.Y.: Modern Chinese paintings: an investment alternative? Southern Economic Journal **59** (1993) 808–816

[31] Pagan, A. and Vella, F.: Diagnostic tests for models based on individual data: a survey. Journal of Applied Econometrics **4** (1989) 29–59

[32] Pesando, J.E.: Art as an investment: the market for modern prints. The American Economic Review **83** (1993) 1075–1089

[33] Pesando, J.E. and Shum, P.M.: The return to Picasso's prints and to traditional financial assets, 1977 to 1996. Journal of Cultural Economics **23** (1999) 183–192

[34] Renneboog, L. and Van Houtte, T.: The monetary appreciation of paintings: from Realism to Magritte. Cambridge Political Economy Society **26** (2002) 331–358

[35] Sacco, P., Santagata, W. and Trimarchi, M.: *L'arte contemporanea italiana nel mondo. Analisi e strumenti*, Collana Opera DARC, Skira, Milano (2005)

[36] Skeels, C.L. and Vella, F.: A Monte Carlo investigation of the sampling behavior of conditional moment tests in Tobit and Probit models. Journal of Econometrics **92** (1999) 275–294

[37] Stein, J.P.: The monetary appreciation of paintings. Journal of Political Economy **85** (1977) 1021–1035

[38] Wieand, K., Donaldson, J. and Quintero, S.: Are real asset prices internationally? Evidence from the art market. Multinational Finance Journal **2** (1998) 167–187

[39] Worthington, A.C. and Higgs, H.: A note on financial risk, return and asset pricing in Australian modern and contemporary art. Journal of Cultural Economics **30** (2006) 73–84

[40] Zanola, R.: The dynamics of art prices: The selection corrected repeat-sales index. Working Paper **85**, Dipartimento di Politiche Pubbliche e Scelte Collettive (POLIS), Università del Piemonte Orientale (2007)

Appendix

Heckit estmation

Table 1: Heckit estimation (1^{st} step)

	dependent variable: y_i				dependent variable: Y_i			
variable	coeff.	s.e.	t-stat	p-value	coeff.	s.e.	t-stat	p-value
constant	-0.0348	0.3226	-0.1078	0.9142	-0.0302	0.3227	-0.0936	0.9254
Characteristics of the artist								
Adami	-0.3406	0.2820	-1.2080	0.2271	-0.3431	0.2820	-1.2164	0.2238
Beecroft	-0.3884	0.5256	-0.7390	0.4599	-0.3935	0.5256	-0.7487	0.4540
Boetti	0.7326	0.2226	3.2916	0.0010***	0.7327	0.2226	3.2913	0.0010***
Burri	0.0056	0.2129	0.0265	0.9788	0.0043	0.2129	0.0204	0.9838
Campigli	0.3642	0.2116	1.7217	0.0851*	0.3651	0.2116	1.7254	0.0845*
Castellani	0.1181	0.2941	0.4015	0.6881	0.1163	0.2942	0.3953	0.6927
Cattelan	0.0390	0.5298	0.0735	0.9414	0.0365	0.5300	0.0689	0.9451
Chia	-0.2070	0.2846	-0.7273	0.4671	-0.2088	0.2846	-0.7336	0.4632
Clemente	-0.4006	0.2953	-1.3566	0.1749	-0.4131	0.2956	-1.3974	0.1623
Cucchi	0.0589	0.3183	0.1849	0.8533	0.0555	0.3184	0.1744	0.8616
Fontana	0.3571	0.1872	1.9070	0.0565*	0.3572	0.1873	1.9073	0.0565*
Kounellis	0.1475	0.3338	0.4418	0.6586	0.1325	0.3347	0.3958	0.6922
Magnelli	0.3987	0.2283	1.7467	0.0807*	0.3979	0.2283	1.7429	0.0813*
Manzoni	0.1895	0.2194	0.8637	0.3878	0.1894	0.2195	0.8630	0.3881
Marini	0.3642	0.2501	1.4561	0.1454	0.3661	0.2501	1.4637	0.1433
Melotti	-0.0838	0.4984	-0.1682	0.8664	-0.0814	0.4985	-0.1633	0.8703
Merz	-0.2495	0.3291	-0.7582	0.4483	-0.2513	0.3292	-0.7634	0.4453
Music	-0.1315	0.2613	-0.5034	0.6147	-0.1354	0.2613	-0.5181	0.6044
Paladino	-0.2377	0.2795	-0.8505	0.3950	-0.2476	0.2798	-0.8848	0.3763
Pomodoro	-1.1805	0.7890	-1.4962	0.1346	-1.1791	0.7895	-1.4935	0.1353
dead	-0.4077	0.1928	-2.1142	0.0345**	-0.4107	0.1929	-2.1288	0.0333**
Physical characteristics								
enamel	-0.6613	0.2568	-2.5752	0.0100**	-0.6560	0.2568	-2.5547	0.0106**
mixed	-0.0662	0.1202	-0.5512	0.5815	-0.0661	0.1202	-0.5499	0.5824
oil	0.0474	0.0921	0.5147	0.6068	0.0493	0.0921	0.5347	0.5928
tempera	0.0299	0.1208	0.2475	0.8046	0.0305	0.1208	0.2521	0.8010
other	0.1384	0.0942	1.4694	0.1417	0.1408	0.0942	1.4956	0.1348
canvas	0.0964	0.0781	1.2350	0.2168	0.0939	0.0781	1.2016	0.2295
paper	-0.0829	0.1009	-0.8217	0.4112	-0.0889	0.1011	-0.8798	0.3790
Artistic characteristics								
authentic	-0.0990	0.1075	-0.9214	0.3569	-0.0972	0.1075	-0.9044	0.3658
catalogue	-0.0161	0.0808	-0.1990	0.8423	-0.0167	0.0808	-0.2064	0.8365
exhibit	0.0185	0.0181	1.0227	0.3065	0.0188	0.0182	1.0341	0.3011
expertise	-0.0663	0.1374	-0.4824	0.6295	-0.0667	0.1374	-0.4852	0.6275
literature	-0.1379	0.0866	-1.5919	0.1114	-0.1369	0.0866	-1.5800	0.1141
owners	0.0152	0.0290	0.5236	0.6005	0.0146	0.0290	0.5023	0.6154
signature	0.0303	0.0670	0.4526	0.6508	0.0301	0.0670	0.4487	0.6536
Sale characteristics								
christies	0.6503	0.1018	6.3914	0.0000***	0.6510	0.1018	6.3977	0.0000***
sothebys	0.8105	0.1011	8.0196	0.0000***	0.8075	0.1011	7.9886	0.0000***
finarte	0.3952	0.1094	3.6134	0.0003***	0.3950	0.1094	3.6120	0.0003***
*d_*1991	0.2600	0.1696	1.5327	0.1254	0.2609	0.1696	1.5384	0.1240
*d_*1992	0.2926	0.1535	1.9063	0.0566*	0.2942	0.1535	1.9161	0.0554*
*d_*1993	-0.3722	0.1522	-2.4454	0.0145**	-0.3716	0.1522	-2.4418	0.0146**

continued on next page

Table 1 — *continued from previous page*

	dependent variable: y_i				dependent variable: Y_i			
variable	coeff.	s.e.	t-stat	p-value	coeff.	s.e.	t-stat	p-value
d_1994	-0.1253	0.1445	-0.8671	0.3859	-0.1445	0.1452	-0.9953	0.3196
d_1995	0.0688	0.1504	0.4573	0.6475	0.0698	0.1504	0.4642	0.6425
d_1996	-0.0857	0.1412	-0.6070	0.5439	-0.0897	0.1414	-0.6344	0.5258
d_1997	-0.4023	0.1452	-2.7701	0.0056***	-0.4006	0.1452	-2.7586	0.0058***
d_1998	-0.1693	0.1530	-1.1062	0.2686	-0.1673	0.1530	-1.0931	0.2743
d_1999	0.1649	0.1392	1.1845	0.2362	0.1657	0.1392	1.1909	0.2337
d_2000	-0.1001	0.1415	-0.7076	0.4792	-0.0983	0.1415	-0.6949	0.4871
d_2001	-0.1726	0.1376	-1.2542	0.2098	-0.1712	0.1376	-1.2440	0.2135
d_2002	-0.2619	0.1353	-1.9349	0.0530*	-0.2605	0.1353	-1.9248	0.0543*
d_2003	-0.0511	0.1391	-0.3670	0.7136	-0.0500	0.1391	-0.3597	0.7191
d_2004	0.4562	0.1519	3.0036	0.0027***	0.4574	0.1519	3.0119	0.0026***
d_2005	0.1862	0.1379	1.3504	0.1769	0.1878	0.1379	1.3623	0.1731
d_2006	0.2148	0.1423	1.5097	0.1311	0.2162	0.1423	1.5197	0.1286

* indicates statistical significance at the 10% level.
** indicates statistical significance at the 5% level.
*** indicates statistical significance at the 1% level.

Table 2: Heckit estimation (2^{nd} step)

	dependent variable: y_i				dependent variable: Y_i			
variable	coeff.	s.e.	t-stat	p-value	coeff.	s.e.	t-stat	p-value
$constant$	3.3029	0.2448	13.4935	0.0000***	3.4341	0.2422	14.1780	0.0000***
$Adami$	-0.7652	0.2167	-3.5311	0.0004***	-0.7534	0.2145	-3.5120	0.0004***
$Beecroft$	-2.2322	0.4070	-5.4853	0.0000***	-2.2017	0.4027	-5.4670	0.0000***
$Boetti$	-0.4771	0.1710	-2.7909	0.0053***	-0.4796	0.1692	-2.8340	0.0046***
$Burri$	1.1525	0.1604	7.1865	0.0000***	1.1472	0.1587	7.2300	0.0000***
$Campigli$	1.2194	0.1589	7.6732	0.0000***	1.2137	0.1573	7.7170	0.0000***
$Castellani$	-0.1641	0.2136	-0.7679	0.4425	-0.1508	0.2114	-0.7130	0.4755
$Cattelan$	0.9315	0.3510	2.6537	0.0080***	0.9186	0.3474	2.6450	0.0082***
$Chia$	-0.5868	0.2137	-2.7464	0.0060***	-0.5754	0.2115	-2.7210	0.0065***
$Clemente$	-0.0013	0.2242	-0.0060	0.9952	0.0025	0.2225	0.0110	0.9910
$Cucchi$	-0.5377	0.2277	-2.3620	0.0182**	-0.5193	0.2253	-2.3050	0.0212**
$Fontana$	1.1076	0.1442	7.6810	0.0000***	1.0934	0.1427	7.6610	0.0000***
$Kounellis$	0.2706	0.2369	1.1422	0.2534	0.2745	0.2353	1.1660	0.2435
$Magnelli$	0.1114	0.1699	0.6556	0.5121	0.1116	0.1682	0.6640	0.5068
$Manzoni$	1.3272	0.1652	8.0327	0.0000***	1.3074	0.1635	7.9980	0.0000***
$Marini$	0.9097	0.1846	4.9268	0.0000***	0.8922	0.1827	4.8820	0.0000***
$Melotti$	-1.0547	0.4120	-2.5602	0.0105**	-1.0569	0.4076	-2.5930	0.0095***
$Merz$	-0.5038	0.2506	-2.0103	0.0444**	-0.4909	0.2480	-1.9800	0.0478**
$Music$	0.3222	0.2004	1.6073	0.1080	0.3218	0.1984	1.6220	0.1048
$Paladino$	-0.5513	0.2105	-2.6192	0.0088***	-0.5430	0.2086	-2.6020	0.0093***
$Pomodoro$	-0.4690	0.7805	-0.6009	0.5479	-0.4601	0.7720	-0.5960	0.5512
$dead$	-0.0305	0.1450	-0.2102	0.8335	-0.0225	0.1437	-0.1570	0.8755
$enamel$	0.1295	0.2290	0.5655	0.5718	0.1303	0.2265	0.5750	0.5652
$mixed$	0.0917	0.0856	1.0718	0.2838	0.0837	0.0847	0.9890	0.3229
oil	0.1565	0.0647	2.4192	0.0156**	0.1508	0.0641	2.3520	0.0187**
$tempera$	0.0706	0.0868	0.8137	0.4158	0.0685	0.0860	0.7970	0.4255
$other$	0.1607	0.0658	2.4430	0.0146**	0.1520	0.0652	2.3330	0.0197**
$canvas$	0.1537	0.0558	2.7551	0.0059***	0.1510	0.0552	2.7340	0.0063***
$paper$	-0.4449	0.0709	-6.2718	0.0000***	-0.4425	0.0705	-6.2800	0.0000***

continued on next page

Table 2 — *continued from previous page*

	dependent variable: y_i				dependent variable: Y_i			
variable	coeff.	s.e.	t-stat	p-value	coeff.	s.e.	t-stat	p-value
$surface$	0.6517	0.0290	22.5074	0.0000***	0.6445	0.0287	22.4870	0.0000***
$squared$	-0.0528	0.0032	-16.5274	0.0000***	-0.0521	0.0032	-16.4930	0.0000***
Artistic characteristics								
$authentic$	0.1071	0.0794	1.3487	0.1775	0.1124	0.0785	1.4310	0.1524
$catalogue$	0.2797	0.0564	4.9577	0.0000***	0.2742	0.0558	4.9120	0.0000***
$exhibit$	0.0266	0.0105	2.5321	0.0113**	0.0258	0.0104	2.4870	0.0129**
$expertise$	-0.2376	0.0939	-2.5317	0.0114**	-0.2311	0.0929	-2.4880	0.0128**
$literature$	0.2850	0.0621	4.5891	0.0000***	0.2848	0.0614	4.6360	0.0000***
$owners$	0.1339	0.0192	6.9852	0.0000***	0.1310	0.0190	6.8990	0.0000***
$signature$	0.0585	0.0464	1.2613	0.2072	0.0562	0.0459	1.2230	0.2213
Sale characteristics								
d_1991	-0.4911	0.1108	-4.4328	0.0000***	-0.4956	0.1097	-4.5200	0.0000***
d_1992	-0.6814	0.1008	-6.7568	0.0000***	-0.6957	0.0998	-6.9680	0.0000***
d_1993	-0.7115	0.1202	-5.9177	0.0000***	-0.6991	0.1190	-5.8760	0.0000***
d_1994	-0.9473	0.1049	-9.0351	0.0000***	-0.9313	0.1051	-8.8590	0.0000***
d_1995	-0.8258	0.1017	-8.1168	0.0000***	-0.8069	0.1007	-8.0150	0.0000***
d_1996	-0.8497	0.0994	-8.5487	0.0000***	-0.8398	0.0987	-8.5100	0.0000***
d_1997	-0.8158	0.1147	-7.1096	0.0000***	-0.7942	0.1135	-6.9950	0.0000***
d_1998	-0.7270	0.1111	-6.5431	0.0000***	-0.7073	0.1099	-6.4340	0.0000***
d_1999	-0.6131	0.0933	-6.5709	0.0000***	-0.5975	0.0924	-6.4690	0.0000***
d_2000	-0.6601	0.1003	-6.5825	0.0000***	-0.6230	0.0992	-6.2790	0.0000***
d_2001	-0.7199	0.0992	-7.2536	0.0000***	-0.6781	0.0982	-6.9060	0.0000***
d_2002	-0.6017	0.1000	-6.0148	0.0000***	-0.5464	0.0990	-5.5210	0.0000***
d_2003	-0.5960	0.0970	-6.1416	0.0000***	-0.5321	0.0960	-5.5420	0.0000***
d_2004	-0.3894	0.0977	-3.9857	0.0001***	-0.3329	0.0967	-3.4410	0.0006***
d_2005	-0.1714	0.0929	-1.8446	0.0651*	-0.1082	0.0920	-1.1770	0.2393
d_2006	0.0564	0.0953	0.5914	0.5542	0.1278	0.0943	1.3550	0.1755
λ_i	-0.5537	0.1733	-3.1942	0.0014***	-0.5504	0.1722	-3.1970	0.0014***

* indicates statistical significance at the 10% level.
** indicates statistical significance at the 5% level.
*** indicates statistical significance at the 1% level.

Table 3. Regression statistics

Dependent variable	y_i	Y_i
Mean of dependent variable	4.0932	4.2402
Std. dev. of dependent variable	1.2911	1.2808
Total observations	2814	2810
Censored observations	803	803
Censored observations (%)	28.5	28.6
Error sum of squares	1078.65	1052.25
S.E. of residuals	0.4371	0.4373
$\hat{\sigma}_{\varepsilon}^2$	0.8231	0.8147
$\hat{\sigma}_{u\varepsilon}$	-0.6727	-0.6755
Akaike Information Criterion	3246.58	3243.88
Bayesian Information Criterion	3573.47	3570.69
Hannan-Quinn Information Criterion	3573.47	3361.82
McFadden R^2 (probit)	0.0685	0.0686
LR test (probit)	230.658	230.674
p-value	0.0000	0.0000
CM test for the normality of u_i	0.7481	0.6789
p-value	0.6879	0.7122
Joint normality test for residuals	157.402	162.969
p-value	0.0000	0.0000

List of variables

Table 4. Characteristics of the artist (N=2817)

variable	description	birth	dead	obs.
Name of the artist				
$Adami$	1 if the author is Valerio Adami, 0 otherwise	1935	-	170
$Beecroft$	1 if the author is Vanessa Beecroft, 0 otherwise	1966	-	9
$Boetti$	1 if the author is Alighiero Boetti, 0 otherwise	1940	1994	212
$Burri$	1 if the author is Alberto Burri, 0 otherwise	1915	1995	126
$Campigli$	1 if the author is Massimo Campigli, 0 otherwise	1895	1971	268
$Castellani$	1 if the author is Enrico Castellani, 0 otherwise	1930	-	114
$Cattelan$	1 if the author is Maurizio Cattelan, 0 otherwise	1960	-	10
$Chia$	1 if the author is Sandro Chia, 0 otherwise	1946	-	155
$Clemente$	1 if the author is Francesco Clemente, 0 otherwise	1952	-	101
$Cucchi$	1 if the author is Enzo Cucchi, 0 otherwise	1950	-	65
$Fontana$	1 if the author is Lucio Fontana, 0 otherwise	1899	1968	720
$Gnoli$	1 if the author is Domenico Gnoli, 0 otherwise	1933	1970	64
$Kounellis$	1 if the author is Jannis Kounellis, 0 otherwise	1936	-	51
$Magnelli$	1 if the author is Alberto Magnelli, 0 otherwise	1888	1971	105
$Manzoni$	1 if the author is Piero Manzoni, 0 otherwise	1934	1963	137
$Marini$	1 if the author is Marino Marini, 0 otherwise	1901	1980	68
$Melotti$	1 if the author is Fausto Melotti, 0 otherwise	1901	1986	8
$Merz$	1 if the author is Mario Merz, 0 otherwise	1925	-	41
$Music$	1 if the author is Zoran Music, 0 otherwise	1909	2005	241
$Paladino$	1 if the author is Mimmo Paladino, 0 otherwise	1948	-	150
$Pomodoro$	1 if the author is Arnaldo Pomodoro, 0 otherwise	1926	-	2
Living status				
$Dead$	1 if the painter is dead at the moment of selling, 0 otherwise			1705
Year of birth				
$Birth$	Year of birth			

Source: Artindex Plus - Gabrius S.p.A.

Table 5. Physical characteristics (N=2817)

variable	description	obs.
Medium		
$collage$	1 if the medium is collage, 0 otherwise	5
$enamel$	1 if the medium is enamel, 0 otherwise	29
$gouache$	1 if the medium is gouache, 0 otherwise	1
$mixed$	1 if the medium is mixed, 0 otherwise	385
$pencil$	1 if the medium is pencil, 0 otherwise	2
oil	1 if the medium is oil, 0 otherwise	1429
$tempera$	1 if the medium is tempera, 0 otherwise	320
$other$	1 if the medium is other, 0 otherwise	1037
Support		
$board$	1 if the support is board, 0 otherwise	185
$canvas$	1 if the support is canvas, 0 otherwise	2254
$cartoon$	1 if the support is cartoon, 0 otherwise	173
$fabric$	1 if the support is fabric, 0 otherwise	75
$marble$	1 if the support is marble, 0 otherwise	5
$masonite$	1 if the support is masonite, 0 otherwise	26
$panel$	1 if the support is panel, 0 otherwise	166
$paper$	1 if the support is paper, 0 otherwise	275
$wood$	1 if the support is wooden base, 0 otherwise	7
$support$	1 if the support is other, 0 otherwise	146
Size		
$surface$	Painting area (in m^2)	
$squared$	Painting squared area	

Source: Artindex Plus - Gabrius S.p.A.

Note: for some paintings different media or different supports are jointly used.

Table 6. Artistic characteristics (N=2817)

variable	description	obs.
$authentic$	1 if the painter has confirmed the authenticity, 0 otherwise	187
$catalogue$	1 if the painting is published on catalogs/monographies, 0 otherwise	680
$date$	1 if the painting is dated, 0 otherwise	1700
$expertise$	1 if the painting is recognised by experts, 0 otherwise	132
$literature$	1 if the painting is cited in literature, 0 otherwise	1049
$signature$	1 if the painting is signed, 0 otherwise	2071
$title$	1 if the painting is titled, 0 otherwise	1722
$exhibit$	Number of exhibitions	
$owners$	Number of previous owners	

Source: Artindex Plus - Gabrius S.p.A.

Table 7: Sale characteristics (N=2817)

variable	description	obs.
Auction houses		
$Curial$	1 if the painting was sold at Art Curial, 0 otherwise	36
$Bonhams$	1 if the painting was sold at Bonhams, 0 otherwise	4
$Bruun$	1 if the painting was sold at Bruun Rasmussen, 0 otherwise	2
$Bukowskis$	1 if the painting was sold at Bukowskis, 0 otherwise	2
$Camels$	1 if the painting was sold at Camels Cohen, 0 otherwise	2
$Christies$	1 if the painting was sold at Christie's, 0 otherwise	914
$Dorotheum$	1 if the painting was sold at Dorotheum, 0 otherwise	9
$Doyle$	1 if the painting was sold at Doyle, 0 otherwise	2
$Finarte$	1 if the painting was sold at Finarte Semenzato, 0 otherwise	536
$Grisebach$	1 if the painting was sold at Grisebach, 0 otherwise	8
$Koller$	1 if the painting was sold at Koller, 0 otherwise	5
$Lempertz$	1 if the painting was sold at Lempertz, 0 otherwise	39
$Neumeister$	1 if the painting was sold at Nuemeister, 0 otherwise	3
$Pandolfini$	1 if the painting was sold at Pandolfini, 0 otherwise	1
$Phillips$	1 if the painting was sold at Phillips, 0 otherwise	37
$Piasa$	1 if the painting was sold at Piasa, 0 otherwise	1
$Porro$	1 if the painting was sold at Porro & C., 0 otherwise	27
$Sothebys$	1 if the painting was sold at Sotheby's, 0 otherwise	1137
$Tajan$	1 if the painting was sold at Tajan, 0 otherwise	43
Marketplace		
$Amsterdam$	1 if the painting was sold in Amsterdam, 0 otherwise	4
NY	1 if the painting was sold in New York, 0 otherwise	363
$Berlin$	1 if the painting was sold in Berlin, 0 otherwise	8
$Paris$	1 if the painting was sold in Paris, 0 otherwise	88
$Cologne$	1 if the painting was sold in Cologne, 0 otherwise	39
$Copenhagen$	1 if the painting was sold in Copenhagen, 0 otherwise	2
$London$	1 if the painting was sold in London, 0 otherwise	1109
LA	1 if the painting was sold in Los Angeles, 0 otherwise	4
$Lugano$	1 if the painting was sold in Lugano, 0 otherwise	17
$Milan$	1 if the painting was sold in Milan, 0 otherwise	994
$Montecarlo$	1 if the painting was sold in Montecarlo, 0 otherwise	3
$Munich$	1 if the painting was sold in Munich, 0 otherwise	3
$Rome$	1 if the painting was sold in Rome, 0 otherwise	140
$Stockholm$	1 if the painting was sold in Stokholm, 0 otherwise	11
$Sidney$	1 if the painting was sold in Sidney, 0 otherwise	1
$Venice$	1 if the painting was sold in Venice, 0 otherwise	17
$Vienna$	1 if the painting was sold in Vienna, 0 otherwise	9
$Zurich$	1 if the painting was sold in Zurich, 0 otherwise	5
Sale date		
d_1990	1 if the painting was sold in 1990, 0 otherwise	242
d_1991	1 if the painting was sold in 1991, 0 otherwise	100
d_1992	1 if the painting was sold in 1992, 0 otherwise	140
d_1993	1 if the painting was sold in 1993, 0 otherwise	109
d_1994	1 if the painting was sold in 1994, 0 otherwise	133
d_1995	1 if the painting was sold in 1995, 0 otherwise	135
d_1996	1 if the painting was sold in 1996, 0 otherwise	148
d_1997	1 if the painting was sold in 1997, 0 otherwise	127
d_1998	1 if the painting was sold in 1998, 0 otherwise	116
d_1999	1 if the painting was sold in 1999, 0 otherwise	190
d_2000	1 if the painting was sold in 2000, 0 otherwise	158
d_2001	1 if the painting was sold in 2001, 0 otherwise	182

continued on next page

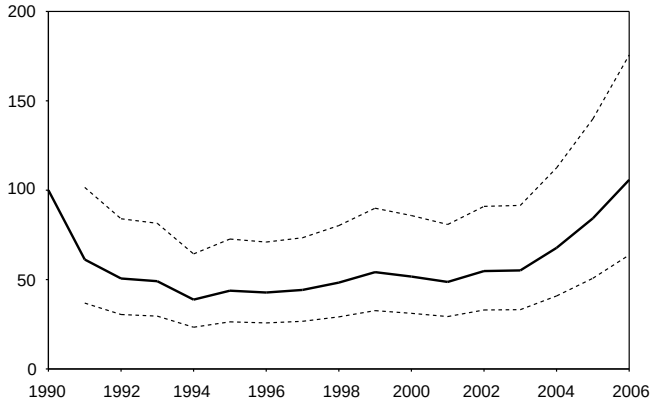
Table 7 — *continued from previous page*

variable	description	obs.
d_2002	1 if the painting was sold in 2002, 0 otherwise	201
d_2003	1 if the painting was sold in 2003, 0 otherwise	206
d_2004	1 if the painting was sold in 2004, 0 otherwise	187
d_2005	1 if the painting was sold in 2005, 0 otherwise	332
d_2006	1 if the painting was sold in 2006, 0 otherwise	211
jan	1 if the painting was sold in January, 0 otherwise	1
feb	1 if the painting was sold in February, 0 otherwise	161
mar	1 if the painting was sold in March, 0 otherwise	245
apr	1 if the painting was sold in April, 0 otherwise	134
may	1 if the painting was sold in May, 0 otherwise	564
jun	1 if the painting was sold in June, 0 otherwise	466
jul	1 if the painting was sold in July, 0 otherwise	38
aug	1 if the painting was sold in August, 0 otherwise	5
sep	1 if the painting was sold in September, 0 otherwise	4
oct	1 if the painting was sold in October, 0 otherwise	433
nov	1 if the painting was sold in November, 0 otherwise	519
dec	1 if the painting was sold in December, 0 otherwise	347
m_i	Pre-sale evaluation (minimum)	2777
M_i	Pre-sale evaluation (maximum)	2777

Fonte: Artindex Plus - Gabrius S.p.A.

Figures

Fig. 1. Price index for the Italian Contemporary Art paintings (I_t)



For each dependent variable, the dotted lines show the index confidence intervals given by $100 \cdot \exp\{\beta_t \pm 1.96 \cdot s.e.(\beta_t)\}$ for the t-th year.

Analysis of the Tourist Sector Investment Appeal Using the PCA in GRETL

Tetyana Kokodey

Sevastopol National Technical University
St. Gogolya 14,
99011 Sevastopol, Ukraine
tkl78@sevcable.net

The article discusses econometrical modeling of a generalized investment appeal indicator for a tourist company and for a tourist sector of a region by the example of Crimea. The latter is further called the rating of investment appeal for a region. The database of the main financial and activity indicators for Crimean tourist companies (2003-2007) was used to calculate the indicators.

Investment policy realization in Ukraine, specifically in the tourist sector, depends on scientifically proved undertakings to attract investments. These undertakings are planned using various econometrical methods to analyze investment appeal of a tourist company.

Table 1 shows data describing the volume of direct foreign investment into the tourist sector of Crimea during 2003-2007yrs.

Table 1. The volume of direct foreign investments into the tourist sector of Crimeaза 2003-2007 yrs. (end of year)

	unit	2003	2004	2005	2006	2007
TOTAL in Crimea	thousand dollars	234045,1	331179,9	460367,0	576971,4	726194,9
The volume of direct foreign investments into the tourist sector of Crimea	thousand dollars.	150312,1	174093,3	217876,7	259879,8	288941,4
Including: Sanatoriums	thousand dollars.	101611,0	112986,6	125497,0	127341,1	143026,0
Per cent to the volume of direct foreign investments	%	67,6	64,9	57,6	49,0	49,5
Tourist companies	thousand dollars.	3908,1	4700,5	37910,5	52495,7	53165,2
Per cent to the volume of direct foreign investments	%	2,6	2,7	17,4	20,2	18,4

The data above shows the increase in the direct foreign investments into Crimea during 2003-2007, nevertheless the relative level stays low as shown on Figure 1.

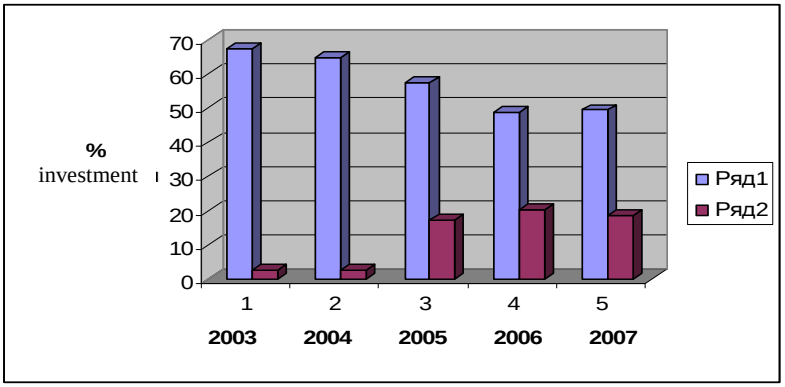
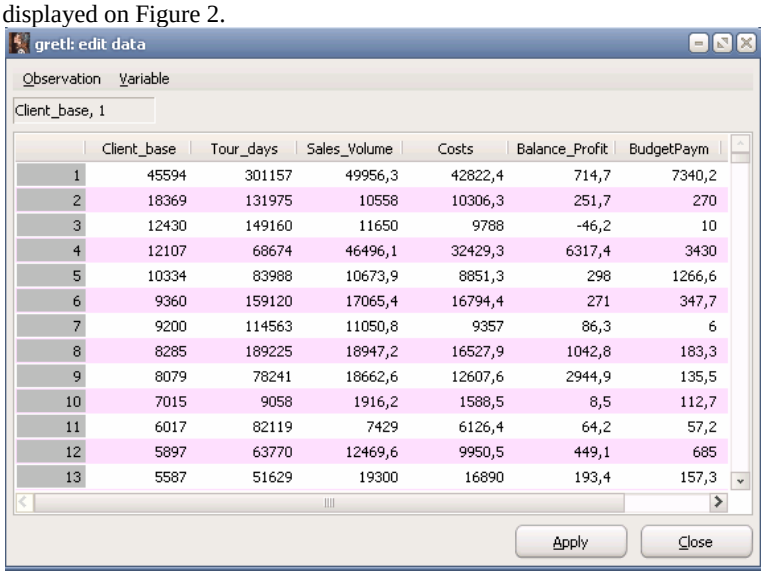


Fig 1. Dynamics of direct foreign investments into the tourist sector of Crimea during 2003-2007 yrs.

The rating of investment appeal for a region, calculated using econometrical methods, can be used as a tool to explain the dynamics presented above, as far as it is a universal instrument of the overall independent assessment of the current state and potential of a region's tourist sector. The rating also provides the key information for an investor regarding both generalized assessment and projections of the operational efficiency of all the region's company.

There are several means to determine the rating as an integral indicator of investment appeal. One of them is GRETL www.sourceforget.net (GNU Regression Econometrics and Time Series Library) open source econometrical software procedures application.

The rating of investment appeal for a region is calculated based on the main financial and activity indicators of all functioning tourist companies of Crimea, using data from 2003 to 2007, fragment of which is displayed on Figure 2.



	Client_base	Tour_days	Sales_Volume	Costs	Balance_Profit	BudgetPaym
1	45594	301157	49956,3	42822,4	714,7	7340,2
2	18369	131975	10558	10306,3	251,7	270
3	12430	149160	11650	9788	-46,2	10
4	12107	68674	46496,1	32429,3	6317,4	3430
5	10334	83988	10673,9	8851,3	298	1266,6
6	9360	159120	17065,4	16794,4	271	347,7
7	9200	114563	11050,8	9357	86,3	6
8	8285	189225	18947,2	16527,9	1042,8	183,3
9	8079	78241	18662,6	12607,6	2944,9	135,5
10	7015	9058	1916,2	1588,5	8,5	112,7
11	6017	82119	7429	6126,4	64,2	57,2
12	5897	63770	12469,6	9950,5	449,1	685
13	5587	51629	19300	16890	193,4	157,3

Fig 2. Fragment of the financial and activity indicators database for Crimean companies in 2007. in Gretl 1.7.1

The following individual indicators were used in calculations, names in parenthesis indicate the corresponding variable name in Gretl:

X_1 - Number of tourists (Client base);
X_2 - Number of tour days (TourDays);
X_3 - Sales Volume (SalesVolume);
X_4 - Balance Profit (BalanceProfit);

X_5 - Budget Payments (BudgetPaym);
X_6 - Costs (Costs).

As far as the indicators $x_1,...,x_6$ are correlated among each other in the substantial extent, the principal components method can be applied to calculate one the most significant principal component y_1 (with the maximum contribution into the overall dispersion of $x_1,...,x_6$) as a linear function of the original indexes, formula (1).

The principal component y_1 can be used as a generalized index of the investment appeal for a company, as far as it contains the majority of information about the company from $x_1,...,x_6$.

$$y_1(x) = w_{11}(\frac{x_1 - \bar{x}_1}{\sigma_1}) + ... + w_{61}(\frac{x_6 - \bar{x}_6}{\sigma_6}); \quad (1)$$

where $\bar{x}_j$ and σ_j — the average and standard deviation of x_j;

w_{j1} – coefficients of the most significant principal component ($\sum_{j=1}^{6} w_{j1}^2 = 1$);

y_1 – the most significant principal component - a generalized index of the investment appeal for a company.

The value λ_1 is the maximum eigenvalue for the first principal component y_1. As far as λ_1 generalizes the majority of observations $x_1,...,x_6$, it can be considered the rating of investment appeal for a region in a given year and further used to track the dynamics of a region's investment appeal.

According to the modeling results obtained in Gretl (Figure 3), the generalized indicator of investment appeal for a company (y_1) for the year 2007 is determined using formula (2):

$$Y1_{2007} = 0{,}452\text{ClientBase} + 0{,}428\textit{TourDays} + 0{,}459\textit{SalesVolume} + 0{,}453\text{Costs} + \\ + 0{,}152\textit{Balance}\,\text{Pr}\,\textit{ofit} + 0{,}416\textit{BudgetPayments} \quad (2)$$

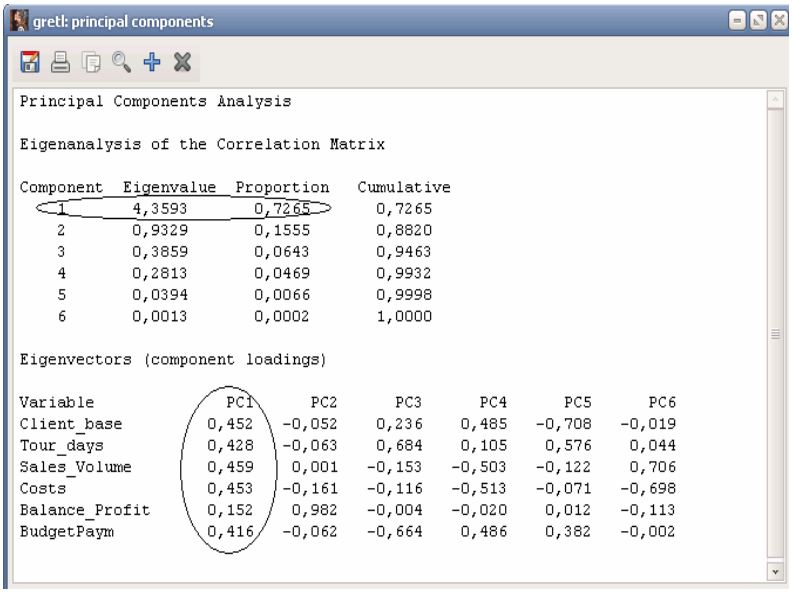


Fig. 3. Modeling results in Gretl using the Principal Components Method for the year 2007

The per cent of initial data $x_1,...,x_6$ embraced by y_1 and included into the formula (2) is 72,65% (Proportion on Figure 3).

Thus, eigenvalue $\lambda_{2007} = 4{,}3593$ can be considered the rating of investment appeal for a region in 2007.

Similar calculations were conducted using 2006 data (Figure 4), formula (3) was obtained:

$$Y1_{2006} = 0{,}457\text{ClientBase} + 0{,}434\textit{TourDays} + 0{,}451\textit{SalesVolume} + 0{,}453\text{Costs} + \\ + 0{,}06\textit{Balance}\,\text{Pr}\,\textit{ofit} + 0{,}436\textit{BudgetPayments} \quad \textbf{(3)}$$

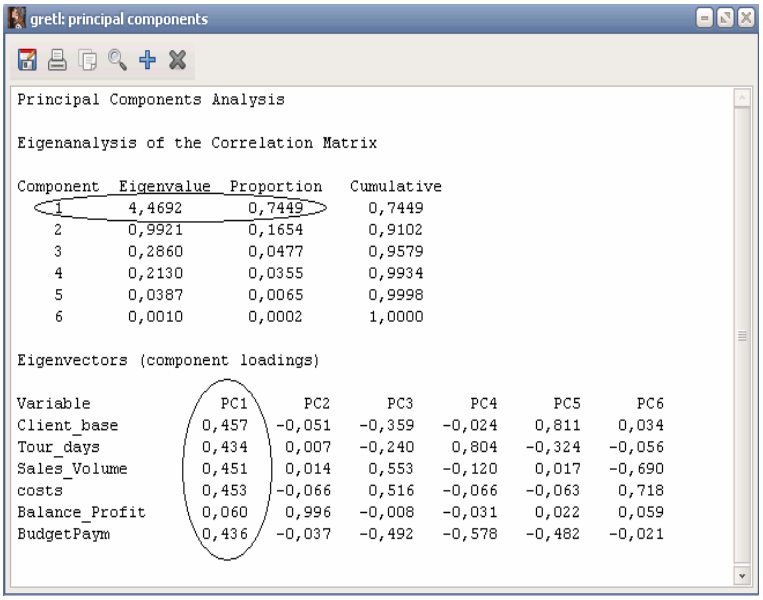


Fig. 4. Modeling results in Gretl using the Principal Components Method for the year 2006

The per cent of initial data $x_1,...,x_6$ embraced by y_1 and included into the formula (3) is 74,49% (Proportion on Figure 4).

Thus, eigenvalue $\lambda_{2006} = 4{,}4692$ can be considered the rating of investment appeal for a region in 2006.

Similar calculations were conducted using 2005 data (Figure 5), formula (4) was obtained:

$$Y1_{2005} = 0{,}447\text{Clien tBase} + 0{,}446\textit{TourDays} + 0{,}472\textit{SalesVolum e} + 0{,}469\text{Costs} - \\ - 0{,}004\textit{Balance}\,\text{Pr}\,\textit{ofit} + 0{,}399\textit{BudgetPaym ents} \quad (4)$$

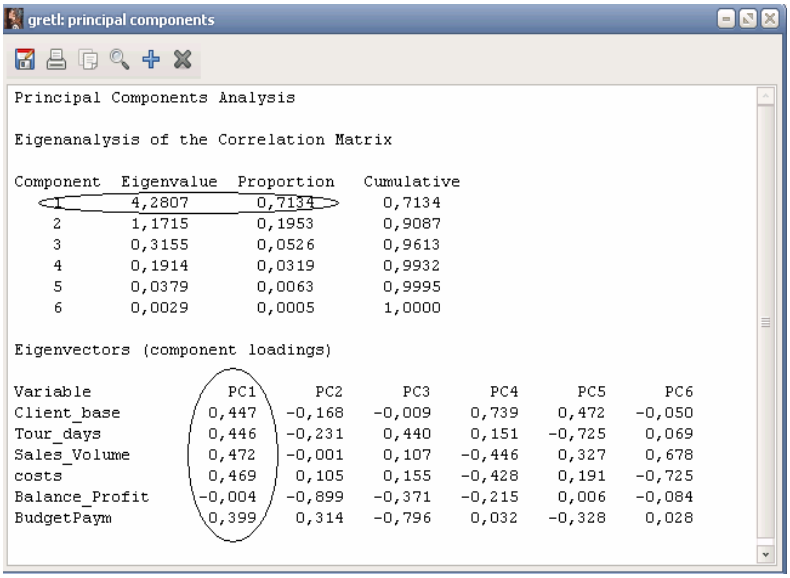


Fig. 5. Modeling results in Gretl using the Principal Components Method for the year 2005

The per cent of initial data $x_1,...,x_6$ embraced by y_1 and included into the formula (4) is 71,34% (Proportion on Figure 5).

Thus, eigenvalue $\lambda_{2005} = 4{,}2807$ can be considered the rating of investment appeal for a region in 2005.

The modeling results above can be summarized in the table 2 below.

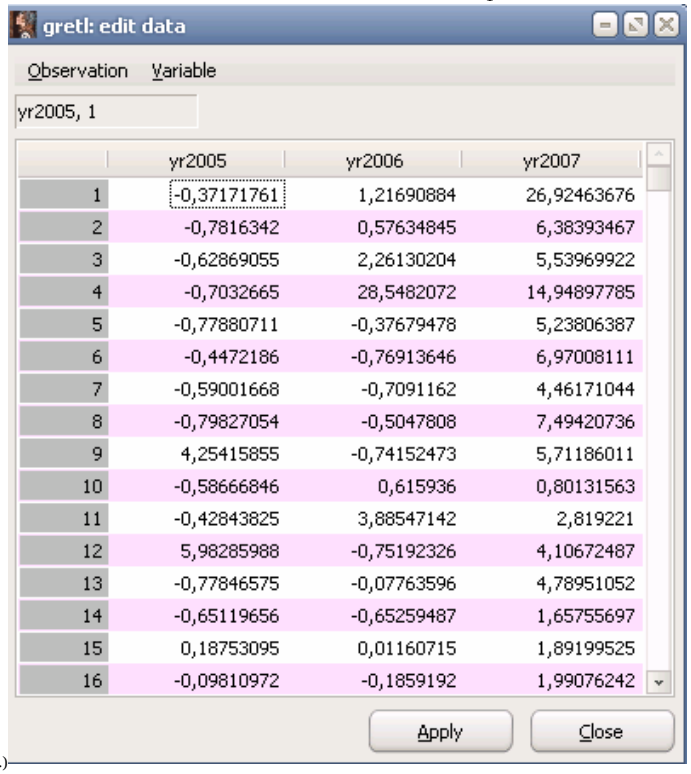
Table 2. Modeling results for the Crimean Region tourist sector investment appeal.

year	The generalized index of the investment appeal for a company, $Y1_i$	Rating of the investment appeal for Crimea, λ_i
2005	$Y1_{2005} = 0{,}447ClientBase + 0{,}446TourDays + 0{,}472SalesVolume + 0{,}469Costs - -0{,}004Balance\,Profit + 0{,}399BudgetPayments$	4,2807
2006	$Y1_{2006} = 0{,}457ClientBase + 0{,}434TourDays + 0{,}451SalesVolume + 0{,}453Costs + +0{,}06BalanceProfit + 0{,}436BudgetPayments$	4,4692
2007	$Y1_{2007} = 0{,}452ClientBase + 0{,}428TourDays + 0{,}459SalesVolume + 0{,}453Costs + +0{,}152Balance\,Profit + 0{,}416BudgetPayments$	4,3593

Fig. 6 shows a fragment of the calculated database of $Y1_{2005}, Y1_{2006}, Y1_{2007}$ for every tourist company in Crimea.

Thus, the principal component method allows for generalization and synthesis of indicial financial and activity indicators $x_1,...,x_6$ for individual companies in a certain year into a generalized indicator of a company's investment appeal y_1

(most significant principal component) and also for calculation of the investment appeal rating for a region's tourist sector λ_1 (eigenvalue for y_1.)



	yr2005	yr2006	yr2007
1	-0,37171761	1,21690884	26,92463676
2	-0,7816342	0,57634845	6,38393467
3	-0,62869055	2,26130204	5,53969922
4	-0,7032665	28,5482072	14,94897785
5	-0,77880711	-0,37679478	5,23806387
6	-0,4472186	-0,76913646	6,97008111
7	-0,59001668	-0,7091162	4,46171044
8	-0,79827054	-0,5047808	7,49420736
9	4,25415855	-0,74152473	5,71186011
10	-0,58666846	0,615936	0,80131563
11	-0,42843825	3,88547142	2,819221
12	5,98285988	-0,75192326	4,10672487
13	-0,77846575	-0,07763596	4,78951052
14	-0,65119656	-0,65259487	1,65755697
15	0,18753095	0,01160715	1,89199525
16	-0,09810972	-0,1859192	1,99076242

Fig 6. A fragment of the calculated database of $Y1_{2005}$, $Y1_{2006}$, $Y1_{2007}$

The same calculation method was used to estimate $\lambda_{2004} = 4{,}01$ and $\lambda_{2003} = 3{,}91$

In order to estimate the dynamics of the calculated ratings in the given time-frame (2003-2007) and to make the projections for the period (2008-2012), the trend analysis method (one of the methods for time series analysis) was applied in Gretl.

According to the modeling results using Ordinary Least Squares method (Figure 7) the trend model of the time series of λ_1 was developed, formula (5):

$$\lambda_1 = 3{,}7985 + 0{,}13578t + \varepsilon, \qquad (5),$$

где t - year;

ε - random error (residuals).

Model (5) was developed based on prior obtained values of $\lambda_{2003}, \lambda_{2004,} \lambda_{2005}, \lambda_{2006,} \lambda_{2007}$.

Prediction of the rating λ_1 for the time frame (2008-2012) is shown on Figure 8.

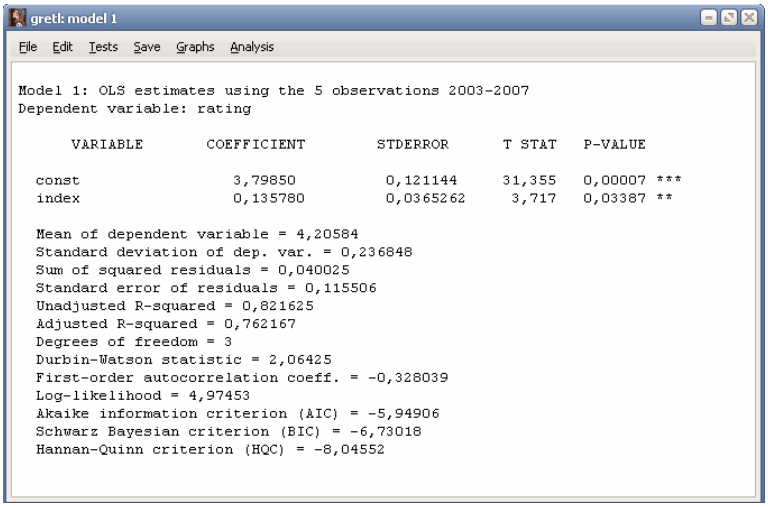
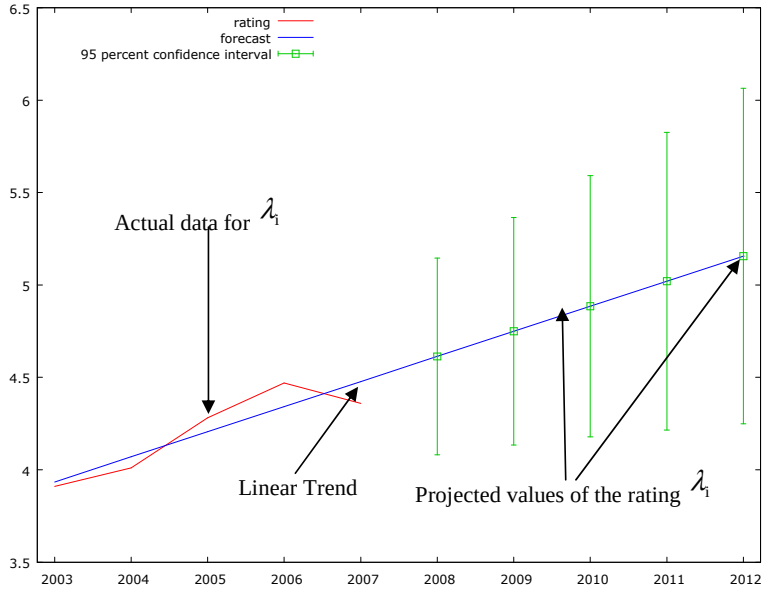


Fig 7. Trend modeling results for λ_i .

The model (5) can be considered adequate and its parameters - valid with a 5% probability of an error.



For 95% confidence intervals, t(3, .025) = 3,182

Obs	Prediction for λ_i (Rating)	std. error	95% confidence interval
2008	4,61318	0,167384	(4,08049, 5,14587)
2009	4,74896	0,193279	(4,13386, 5,36406)
2010	4,88474	0,222181	(4,17766, 5,59182)
2011	5,02052	0,253061	(4,21517, 5,82587)
2012	5,15630	0,285279	(4,24841, 6,06419)

Fig 8. Dynamics and projection for λ_i (Crimea investment appeal rating)

Overall positive dynamics of the investment appeal rating for Crimea can be indicated during the period 2003-2007, (Figure 8). The rating value is going to reach 4,61318 in 20008 and 5,15630 in 2012 according to the projections using the linear trend model (5).

Modeling results obtained in the research allow to make a conclusion that investment appeal of Crimea is going to increase 5,82% in 2008 and 18,28% in 2012 due to the positive dynamics of individual financial and activity indicators of tourist companies in the region.

Has the European Structural Fisheries Policy Influenced on the Second Hand Market of Fishing Vessels[1]?

Ikerne del Valle[*], Kepa Astorkiza,[*] and Inmaculada Astorkiza[*2]
[*] Department of Applied Economics V. University of The Basque Country (UPV-EHU)
Avda. Lehendakari Agirre No. 83, 48015 Bilbao, Spain
ikerne.delvalle@ehu.es

Abstract. The main objective of this paper is to analyze the potential effects that the principal instrument of European structural fisheries policy, that is the Multi-Annual Guidance Plans (MAGP), may have exercised in the second hand market of fishing vessels. In order to test the starting hypothesis that the shortage generated in the periods with the most significant and radical capacity adjustment may have influenced increases on prices of vessels, the main determinants of the second hand market of fishing vessels are analyzed by estimating and discriminating among alternative hedonic price models, such as the linear, log-linear, semi-log, mixed semi-log, TRANSLOG or Box-Cox. Especial attention is paid not only to the correct specification of the numerical variables by carrying out a battery of tests to discriminate among alternative non-nested models, but also to checking for structural change and the convenience of using alternative estimation methods including OLS, weighted least squares (WLS) and trimmed least squares (TLS), based on heteroskedasticity and OLS regression diagnostic. Pooled data is available for 228 transactions taking place during the period 1985-2005 for the Basque artisan and trawling fleets. The principal results achieved indicate that the higher increase of prices precisely happens under the MAGP with major capacity adjustment.

Keywords: European Fisheries Policy, Second Hand Market of Fishing Vessels, Hedonic Model.

---

1 This study has received financial from the EFIMAS project (No SSP8-CT-2003 502516) and ETORTEK2003/IMPRES (Basque Government).

2 The authors are grateful to conference participants at 4th Conference Developments in Economic Theory and Policy (10-11 July 2008, Bilbao) and 35th Annual Conference of the Eastern Economic Association (Feb27-March1 2009, New York). All errors and opinions are the author's responsibility.

1 Background and Problem Definition

Since the mid eighties the European Fisheries Structural Policy has been leaning on the figure of Multi Annual Guidance Plans (MAGP) as the principal vehicle to adapt fleet's fishing capacity to the abundance of the fishing stocks. Basically the design of MAGP has been oriented to the reduction of fishing effort by the limitation of capacity in order to adjust fleets to the real biological situation of fish stocks and also to incentive the modernisation and competitiveness of fishing fleets. In practice the principal materialisation of the measure has been the elimination of vessels by giving incentives in the form of decommissioning grants, the exportation of the vessel to third countries outside the European Union, and even, although is less degree, the reallocation of vessels to another activities. Each MAGP establishes the specific fleet reduction target by member States for a period of 5 years starting from 1982.

Our initial hypothesis is that the shortage generated in the periods with the most significant and radical capacity adjustment, may have influenced increases on prices of the second hand market of vessels. Notice that since entry to the fisheries is limited to those having an authorised vessel, the only way for a firm to add another vessel to the business is by accessing to the second hand market. Moreover, since the effective adjustment patterns differ by State and fishing sub-sector, one may expect disparities on the influence on prices depending not only on MAGP but also on the State and fishing sub-sector. Therefore, some of the related questions we intend to answer are: Have MAGP influenced on the prices of second hand market of fishing vessels? Have they affected homogeneously to the different fishing sub-sectors? Is this potential effect on prices especially significant during some of the MAGP? Is there any interaction effect between the previous the fishing sub-sector and MAGP? We focus on the Basque artisan and trawling fleets to address the researching questions set out.

Vessels constitute a heterogeneous asset. Even when the specifications and their technical characteristics are the same, if the age differs, the degree of deterioration differs accordingly, so that one might say that there are no two identical vessels, or to put in another words, vessels have the particularity of few equivalents. Moreover, the quality of vessels may change with time owing to technological progress. Heterogeneity and potential quality improvements justify the need of methods able to compare heterogeneous vessels prices at different times. These methods should distinguish between movements in prices and changes in the composition of vessels sold from one period to the next. Moreover the objectives of the paper demand a modelling approach open to accommodating for structural changes. The structure of the market changes in accordance with the changes in the preferences of vessels owners, which are influenced in turn by the situation of the fisheries, and

specifically by the fisheries policy. This implies that in more or less degree but the model should be open to structural flexibility.

There are two price modelling approaches that specifically focus on the issues resulting from heterogeneity and changes in quality: the hedonic approach ([1], [2], [3]) and the repeat sales technique [4]. The former controls for quality by using regression models with the attributes of the good as independent variables, while in the second quality control is directly achieved by only including the transactions involving the same assets. The hedonic approach as three principal advantages. First, it uses all of the information on sales in each sample period and not just the data that can be matched, which mitigates the sample selection bias issue. Second, it can adjust for the effects of depreciation if the age of the structure is known at the time of sale. Third, it can adjust for the effects of renovations and repairs if expenditures on renovation and extensions are known at the time of sale.

Different basis explain our methodological choice in favour of the hedonic approach. For one side, since the fluidity of the second hand fishing vessel market is rather low, the sample selection bias would be an extremely large problem if the repeated sales method was adopted. Moreover, we could not get a usable database containing repeated sales transactions of the same vessels because there were an insufficient number of matched observations. Last but not least, since the Boskin report [5], hedonic analysis is gaining academic acceptance as a tool for quality adjustment. Proof of this is it has been applied in a great variety of goods and services, such as automobiles ([2], [6], [7]), Internet services ([8], [9]) and very especially computers ([10], [11]) and houses ([12], [13], [14], [15], [16]), by far the two most popular products for which much of the hedonic research has been concentrated on. To our knowledge this is the first application following the hedonic approach that concerns fishing vessels.

Two alternative approaches have been followed in the literature when dealing to incorporate structural change in the hedonic regression framework ([16], [14], [17]): the characteristic price index approach (CPI) ([2], [6]) and the time dummy variable method (TDV). In the CPI a separate regression is performed for each time period included in the sample, which allows both the coefficients of the good's characteristics and the intercept to change across different periods. The TDV approach consists of performing a pooled regression including time dummy variables to discriminate among different time periods, and thus estimating a common set of coefficients for the attributes, often under the assumption of time independent implicit prices. Compared to the TDV, the CPI has the advantage of being completely structurally unrestricted. Notice however that additional flexibility might be gained in the TDV approach by estimating implicit prices with a time trend, including a series of time dummies for each hedonic characteristic or by including cross terms between the hedonic characteristics and time dummies.

Although the CPI is clearly preferable due to its less demanding assumptions and its structural flexibility, not only the own objectives of the paper but also the data structure and availability (in some individual years we have not enough data to undertake the regression analysis) entails the adoption of the TDV. Thus in the framework of the time dummy variable method, we follow the approach developed by [18]), [19], [20]) and [15] in order to face the structural change and separate the data in different sub-periods using exogenously determined breaking points based on the different MAGP periods. The breaking points of structural change may also be endogenously identified using a switching regression model [15]. However, for the purposes of our study it is reasonable to divide the entire sample into 5 sub-samples, namely "$MAGP_1$", "$MAGP_2$", "MAGP3", "$MAGP_4$" and SUB, attending to the MAGP under which the transaction was materialized. In order to gain the additional structural flexibility needed to test whether each plan has affect in different way to vessels belonging to different sub-sectors, artisan or inshore vessels (B) and offshore or trawler vessels (A) the interaction term "$MAGP_i$ * $SECTOR_i$" will be also included in the model.

2 Data and Fleet Performance

Sample data is available for N=228 transactions of the second hand market of fishing vessels taking place in the Basque Country during the period 1985-2005. N_B=159 transactions correspond to artisan vessels (B), and the remaining N_A=69 are trawlers (A). The information includes the transaction price (P) (in € 2006), gross tonnes of the vessel (GT), horse power of the engine when the sale took place (HP), length (ESL), the AGE of the vessel, the sector the vessel belongs to (i.e. artisan (B) or offshore vessels (trawlers) (A) and the date of the transaction. Following the time dummy variable method we are dividing the data using the breaking points determined by each MAGP. Thus, the category $MAGP_1$ groups all the transactions taking place during the period (1982-86), $MAGP_2$ the ones during the validity of the second plan (1987-1991), $MAGP_3$ includes the ones during the years (1992-1996), MAGP4 the transactions happening during (1997-2001) and finally SUB the ones in the period (2002-2005). Data sample statistics are summarised in Table 1.

Table 1. Data Sample Statistics.

	Variable	Mean	Median	ST	CV	N_{TI}	%
Artisanal vassels (B)	P_B	286.080	183.170	329.410	1,15	-	-
	GT_B	50,09	24,78	48,64	0,97	-	-
	ESL_B	18,42	16,20	7,46	0,4	-	-
	AGE_B	18,38	19,00	8,23	0,44	-	-
	HP_B	242,77	170,00	201,33	0,79	-	-
	$PMAGP_{1B}$	145.705	52.551	188.996	1,29	15	9,43
	$PMAGP_{2B}$	292.127	190.367	246.597	0,84	24	15,09
	$PMAGP_{3B}$	269.879	182.762	324.574	1,20	44	27,67
	$PMAGP_{3B}$	310840	200.573	353.962	1,13	55	34,59
	$PSUB_B$	348.561	218.672	328.855	0,94	21	13,21
Trawlers (A)	P_A	1.248.000	1.057.700	993.700	0,79	-	-
	GT_A	239,36	230,81	100,94	0,42	-	-
	ESL_A	34,32	33,60	6,61	0,19	-	-
	AGE_A	20,12	20,00	6,85	0,34	-	-
	HP_A	772,04	750,00	308,89	0,4	-	-
	$PMAGP_{1A}$	343.331	364.944	175.509	0,51	9	13,04
	$PMAGP_{2A}$	697.471	910.969	467.175	0,66	11	15,94
	$PMAGP_{3A}$	1.457.085	1.293.103	799.756	0,54	26	37,68
	$PMAGP_{3A}$	1.190.145	1.085.686	542.831	0,45	12	17,39
	$PSUB_A$	2.107.555	1.706.94	1.596.29	0,75	11	15,94

N_B=159, N_A=69 N=226
Source: Public Record

Figure 1 and Table 2 show how the different MAGPs have affected the Basque artisan and offshore fleets obeying alternative indicators such as the number of vessels (NB), the total gross tonnage or capacity (GT), total horsepower (HP) and the number of fishermen (L). First, it seems that the $MAGP_1$ hardly did affect the Basque fleet. For one side, since Spain didn't join the European Economic Community (EEC) until 1986, Spanish fleets were not under the requirements of the $MAGP_1$. For another, the level of compliment of European fleets was really low and the deviations above the targets were remarkable even for the member States. In fact, at the end of the $MAGP_1$ the capacity and power increased 4.5% and 8.1% respectively. Second there is an asymmetric path in the degree and temporal affectation on each of the fleets. Contrary to what happened in the artisan fleet, the offshore fleet had an imperceptible adjustment during $MAGP_2$. Even the number of vessels (NB_A), the capacity (GT_A), the power (HP_A) and also the direct employment (L_A) increased during $MAGP_2$. However, in the $MAGP_3$ the negative variations in the indicators fluctuate between 25% and 45%. This negative pattern continues across $MAGP_4$, although much more slightly. This really means that the effective restructuring of the offshore sub-fleet took

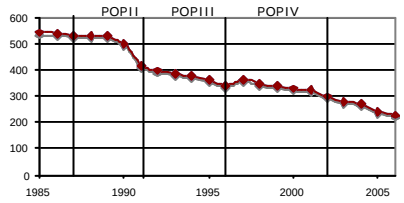
place during $MAGP_3$. In the case of the artisan fleet, the adjustment has been more gradual and uniform. However, it seems that the $MAGP_3$, and very specially the $MAGP_2$ are the ones implying the most dramatic and immediate adaptations. Notice that practically the totality of the adjustment during $MAGP_2$ took place in hardly a couple of years (1989 and 1990). Although the reduction on NB_B is sensibly higher during $MAGP_2$, the fact that the variation on GT_B and HP_B was considerably lower may indicate that it were the smallest vessels the ones, which were withdrawal.

Thus taking into account that the adjustment paths have been different attending to each of the sub-fleets and MAGP, it seems reasonable to cross the two dummy variables (ie. $SECTOR_i$ and $MAGP_i$). This way all the transactions will be grouped in 10 categories. The inshore or artisan vessels sold during $MAGP_1$ have been established as the base category ($MAGP_{1B}$). $MAGP_{1A}$ corresponds to the offshore vessels sold during $MAGP_2$, and so on until completing the rest of the subgroups (i.e. $MAGP_{2B}$, $MAGP_{3A}$, $MAGP_{3B}$, $MAGP_{4A}$, $MAGP_{4B}$, SUB_A y SUB_B.

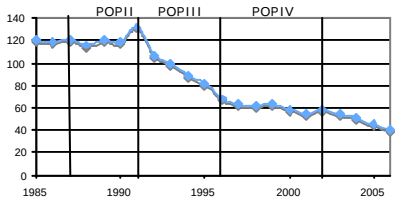
Table 2. Variation (%) on NB, GT, HP and L by sub-sector and MAGP.

MAGP	Period	$\%NB_B$	$\%NB_A$	$\%GT_B$	$\%GT_A$	$\%HP_B$	$\%HP_A$	$\%L_B$	$\%L_A$
$MAGP_1$	1982-86	-	-	-	-	-	-	-	-
$MAGP_2$	1987-91	-21,35	9,09	-9,90	15	-9,13	13,38	-8,48	-0,14
$MAGP_3$	1992-96	-13,75	-36,45	-15,29	-35,93	-8,28	-26,28	-25,43	-41,38
$MAGP_4$	1997-01	-18,03	-7,94	-10,08	-10,36	-13,13	-14,61	-18,97	-15,73
SUB	2002-06	-18,44	-25,93	-17,70	-29,82	-16,96	-33,34	-	-

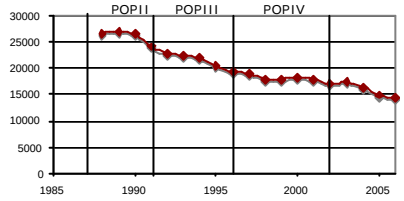
Figure 1. Evolution of NB, GT, HP and L across MAGP and Fishing Sector.



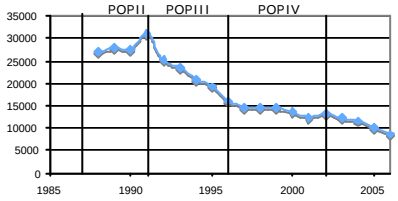
Number of artisan vessels (NB_B)



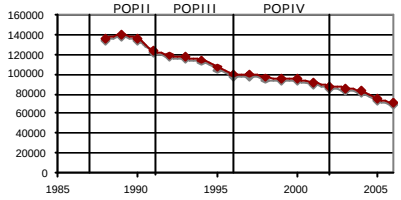
Number of trawlers (NB_A)



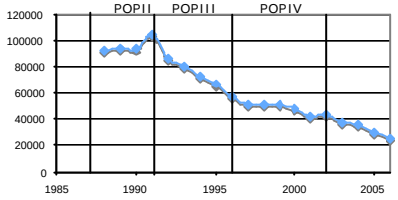
GT artisan fleet (GT_B)



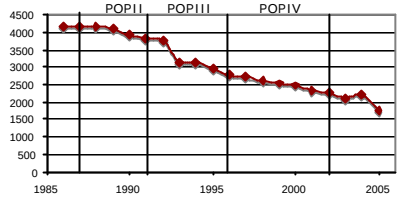
GT trawlers (TRB_A)



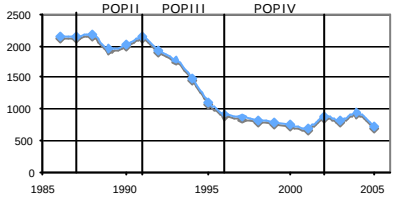
HP artisan fleet (HP_B)



HP trawlers (HP_A)



Employment artisan fleet (L_B)



Employment trawlers (L_A)

3 An Empirical Model for the 2nd Hand of Fishing Vessels

Following the hedonic approach, vessels are considered to be composite products formed by a heterogeneous set of attributes, whose market price at time t (P_{it}) is determined by a set of characteristics $X=(X_1, X_2,\ldots, X_n)$ (i.e. GT, AGE, the fishing SECTOR the vessel belongs to) and by the potential effects that the European Fisheries Structural Policy, via MAGP, may have exercised on the second hand market ($MAGP_t$). Therefore, when acquiring a vessel, we can consider the price to be the sum of the price paid for each one of its attributes [(X_i=GT, AGE, SECTOR), $MAGP_t$)], so that an implicit price, or hedonic price exists for each one of the attributes defining the vessel.

Hence, assuming an optimal behaviour by buyers and sellers the second hand market fishing vessels price hedonic function may be estimated under market equilibrium and using transactions data P_{it}=f((X_i=GT, AGE, SECTOR), $MAGP_t$). The essence of the approach consists of finding what portion of the price is determined by each of the atributtes (X_i, $MAGP_t$). This information is obtained by calculating the partial derivative of the price with respect to each of the attributes ($\partial P_{it}/\partial X_i, \partial P_{it}/\partial MAGP_t$). Besides the marginal willingness to pay for an additional unit of each of the vessels' characteristics, this method also gives us the *structural policy asset*, allowing us to obtain an estimate of its monetary value. To put in another words, we focus on vessels of the same quality (i.e. vessels sharing identical characteristics (X_{it}), and then compare the hedonic price at different time periods grouped in each of the MAGP. This way, the part of overall price change from one MAGP to another, which is not accounted by the changes in characteristics may be then interpreted as pure price change related to changes in economic policy. If this price change is significantly great under the most radical fleet adjustment one may derive that the structural policy is affecting the market.

The modelling requirements implicit in the applied hedonic regression literature focuses the attention on three major points: a) the specification of the functional form; b) the configuration of models suitable to test for structural break; and c) the discussion of the appropriate estimation method, mainly discriminating among OLS and weighted least squares (WLS). All of these topics are covered in the next three subsections. Since this is the first application of the hedonic framework to fishing vessels, special attention will be plaid to the functional form selection procedure. This procedure is based on two steps. First the performance of a priori established functional forms in undertaken attending to the usual goodness of fit measures and the signs of the estimated parameters. Second a battery of non-nested test is completed in

order to check the performance of alternative transformations of the numerical variables included in the model chosen in the first step. Afterwards the issue of the structural time stability is addressed in the framework of the previously selected function by estimating an augmented model incorporating time varying slopes for the numerical independent variables; and carrying on a related F test. Finally the convenience of using other estimation methods instead of OLS is considered based on heteroskedasticity tests and also the regression diagnostic. The estimated coefficients derived by OLS, WLS and trimmed least squares (TLS) are compared in other to see if the results do not radically change depending on the estimation method.

3.1 The Functional Form

Although the specification of the functional form of hedonic regression models has been the subject of considerable debate, the hedonic theory does not give an explicit answer to this central issue and there is no a priori structural restriction on the choice of functional forms. The general view is that the functional form of a hedonic model is purely an empirical issue (Rosen, 1974) and that therefore decisions should be made on a case-by-case basis, taking into account several interrelated issues such as, the purpose for which the hedonic regression is undertaken, the kind of data used, the estimation method, the appropriateness of the derived empirical results and also the results of homoskedasticity tests.

In most empirical studies carried on durable goods using the hedonic framework the model selection procedure is limited to choosing among three popular functional forms (the linear (1), the semi-log (2) and the log linear (or double log) (3)) based on the usual goodness of fit measures (R^2, the standard error of the regression, AIC, etc.). Notice that strictly speaking when dummy variables are included, since these are not transformed, equation (2) is in fact a mixed functional form. Besides, all of the variables of the right side of the equation need not have the same form. For example Barzyk [21] employed a mixed functional form in which some of the variables appear linearly and some logarithmically depending on the data fit and the empirical standards summarized in (Wooldridge [22])[3]. Following the terminology in Triplett (2004) we will refer to these forms as mixed–semilog functions. The semilog and log-linear forms are usually preferred, mainly because they generate a better goodness of fit and because heteroskedasticity is mitigated [23]. However, a simple log model may not be the correct specification due to the

[3] When a priori choosing which of the variables are transformed it is usual to take logs when working with monetary variables taking high values. However, the variables measured in years normally appear in their original form.

possible nonlinear relationship between price and characteristics and interactions between characteristics. Hence joint with the above-mentioned forms, the mixed semilog with quadratic and cross terms (4), the mixed semilog with quadratic but not cross terms (4^R) and the translog (5) will be also tested. Furthermore, attention will be plaid on the Box-Cox model (Box and Cox, 1964) (6), which nests the linear, the semilog and log-linear forms.

Thus the functional forms a priori considered in this paper (i.e. the linear, semi-log, double log, mixed-semilog, translog, Box-Cox) are listed as follows:

$$P_i = b_{0t} + \sum_{k=1}^{K} b_k X_{ikt} + \sum_{s=2}^{t} d_s MAGP_s + e_{it} \quad (1)$$

$$\ln P_{it} = b_{0t} + \sum_{k=1}^{K} b_k X_{ikt} + \sum_{s=2}^{t} d_s MAGP_s + e_{it} \quad (2)$$

$$\ln P_{it} = b_{0ti} + \sum_{k=1}^{K} b_k LnX_{ikt} + \sum_{s=2}^{t} d_s MAGP_s + e_{it} \quad (3)$$

$$\ln P_{it} = b_{0t} + \sum_{l=1}^{L} b_l LnX_{ilt} \sum_{l=1}^{L} b_l (LnX_{ilt})^2 + \sum_{k=1}^{K} b_k X_{ikt} + \sum_{k=1}^{K} b_k (X_{ikt})^2 + \sum_{s=2}^{t} d_s MAGP_s + e_{it} \quad (4)$$

$$\ln P_{it} = b_{0t} + \sum_{k=1}^{K} b_k LnX_{ikt} + 1/2 \sum_{i=1}^{k} \sum_{j=1}^{k} b_{ij} LnX_{it} Ln\, X_{jt} + \sum_{s=2}^{t} d_s MAGP_s + e_{it} \quad (5)$$

$$P_{it}^{\,l} = b_{0t} + \sum_{k=1}^{K} b_k MAGP_{ikt}$$

$$P_{it}^{\,l} = \begin{cases} \dfrac{P_{it}^{\,l} - 1}{l} & if \quad l \neq 0 \\ LnP_{it} & if \quad l = 0 \end{cases} \quad (6)$$

where P_{it} denotes the prices of a vessel (for all the periods included in the sample), $ß_k$ is the coefficient of characteristic k, X_{ikt} the value of characteristic k of vessel i in period t, $ḍ$ the parameter of the time dummy variable in MAGPs and e_{it} the random disturbance term. $MAGP_s$ is the time dummy variable, and takes a value of 1 if the transaction occurs at the certain period t, and 0 otherwise. This model configuration is said to be a structurally restricted hedonic model (SRHM) because it assumes that the regression coefficient $ß_k$ of the vessel-price determining factor X_{ikt} is constant throughout all the periods. However in the next subsection we are adding structural flexibility by allowing one of the attributes (i.e. SECTOR) to change throughout all the periods.

As well as the signs and value of the coefficients and the usual measures (i.e. adjusted R^2, the standard error of the regressions, F test, Akaike Information Criterion (AIC), Schwarz Bayesian Criterion (SBC) and Hannan-Quinn criterion (HQC), the criteria for comparing among alternative model

specification includes a full battery of nested on non-nested tests: Ramsey's RESET test, Davidson-MacKinnon (1981) (DM) test, Minzon y Richard (1986) (MR) test, Wooldridge (1994) and Box Cox test (1964). Notice that while RESET is limited to nested models, DM `F` and MR `t` tests are useful to discriminate between non-nested models of the independent variables, provided that the model includes the same dependent. However DM and MR are not valid in order to contrast among alternative non-nested models when they have different dependent variables, such as $\log(P_i)$ against P_i. Since in these cases you cannot make direct comparisons of R^2 or the sums of the squares of the residuals (SSR), additional test, such the Box-Cox test (1964) (BC) and Wooldridge (1994) tests (W) are required. The OLS results for the lineal (1), semi-log (2), log-linear (3), mixed semilog with quadratic and cross term between TRB and AGE (4), mixed-semilog with quadratic but no interaction term (4^R), and the translog (5) are summarised in Table 3. For one side, attending both to adjusted R^2, the AIC, SBC and HQC the linear model performs worse than the others, although at least it passes the RESET test for model configuration. The models including quadratic and or cross terms perform better attending to the R^2, the AIC, SBC and HQC. Besides, all of them pass the RESET. Moreover apart from the translog the signs for the estimated parameters are the expected ones. Both the AIC and the rest of the models selection criteria and F test for the omission of the cross term between AGE and TRB constitute complement arguments in favour of model (4^R) [F(1, 213)=0.05, p-value=0.8147].

Table 3. Regression results for alternative functional forms

Variable	Linear(1)	SemiLog(2)	Loglinear(3)	MSemilog^{T}(4)	MSemilog(b)^{T}(4)^{R}	Traslog(5)
	165,017	7.6785	8.3568	6.7097	6.7621	6.0342
constant	(145,761)	(0.2829)***	(0.3563)***	(0.4768)***	(0.4204)***	(0.8550)***
	[93,286]*	[0.2903]***	[0.3537]***	[0.4282]***	[0.3604]***	[0.8950]***
	3034.09	1.0865	1.0824	2.0662	2.0560	2.00235
GT	(481.10)***	(0.0525)***	(0.0528)***	(0.2250)***	(0.2203)***	(0.2628)***
	[577.78]***	[0,0530]***	(0.0520)***	[0.1960]***	[0.1845]***	[0.2515]***
	-12810.9	-0.0334	-0.4509	(-0.0777)	-0.0813	0.555934
AGE	(4169.69)***	(0.0069)***	(0.1016)***	(0.0309)**	(0.0268)***	(0.5687)
	[4047.49]***	[0.0069]***	[0.1038]***	[0.0299]***	[0.0251]***	[0.5739]
	-	-	-	-0.1494	-0.1507	-0.145207
GT^2				(0.0339) ***	(0.0268)***	(0.0337)***
				[0.0291]***	[0.0299]***	[0.0294]***
	-	-	-	0.0011	0.0011	-0.223959
AGE^2				(0.0006)*	(0.00067)*	(0.1161)*
				[0.00062]*	[0.00062]*	[0.1137]*
	-	-	-	-0.0011		0.00620675
GT*AGE				(0.0047)	_	(0.0627)
				[0.0046]		[0.0612]
	-266,610	-0.3497	-0.3084	0.0960	0.1058	0.0725587
$MAGP_{1A}$	(219,151)	(0.3507)	(0.3526)	(0.3532)	(0.3500)	(0.3523)
	[116,470]**	[0.3163]	[0.3155]	[0.3041]	[0.3075]	[0.3091]
	169,240	0.5414	0.5069	0.6315	0.6272	0.633033
$MAGP_{2B}$	(162,127)	(0.2642)**	(0.2656)*	(0.2548)**	(0.2536)**	(0.2561)**
	[79,416]**	[0.3163]	[0.3418]	[0.3415]*	[0.3359]*	[0.3429]*
	117,610	0.1497	0.1410	0.6029	0.6094	0.600905
$MAGP_{2A}$	(207,621)	(0.3303)	(0.3327)	(0.3333)*	(0.3314)*	(0.3337)*
	[161,793]	[0.4046]	[0.4104]	[0.4408]	[0.4394]	[0.4423]
	200,915	1.1809	1.1385	1.1671	1.1668	1.1912
$MAGP_{3B}$	(147,531)	(0.2412)***	(0.2422)***	(0.2316)***	0.2311***	(0.2329)***
	[63,774]***	[0.2683]***	[0.2683]***	[0.2644]***	[0.2647]***	[0.2676]***
	725,644	0.9941	0.9771	1.6162	1.6182	1.59677
$MAGP_{3A}$	(190,004)***	(0.2783)***	(0.2804)***	(0.3006)***	(0.2998)***	(0.3005)***
	[182,417]***	[0.2888]***	[0.2864]***	[0.3082]***	[0.3086]***	[0.3106]***
	261,139	1.3240	1.2924	1.3060	1.3071	1.32225
$MAGP_{4B}$	(144,648)*	(0.2362)***	(0.2371)***	(0.2267)***	(0.2261)***	(0.2279)***
	[64,041]***	[0.2502]***	[0.2470]***	[0.2436]***	[0.2431]***	[0.2517]***
	606,392	1.1692	1.1160	1.7333	1.7253	1.72128
$MAGP_{4A}$	(213,315)***	(0.3272)***	(0.3225)***	(0.3393)***	(0.3368)***	(0.3397)***
	[199,991]***	[0.3052]***	[0.2470]***	[0.3011]***	[0.2948]***	[0.3025]***
	231,166	1.3979	1.3468	1.3036	1.3046	1.3598
SUB_B	(166,266)	(0.2709)***	(0.2725)***	(0.2633)***	(0.2627)***	(0.2615)***
	[69,828]***	[0.2709]***	[0.2732]***	[0.2675]***	[0.2675]***	[0.2743]***
	1,553,830	1.4950	1.4742	1.987	1.9834	1.97449
SUB_A	(210,933)***	(0.3313)***	(0.3334)***	(0.3363)***	(0.3352)***	(0.3370)***
	[541,888]***	[0.3770]***	[0.3814]***	[0.3810]	[0.3795]***	[0.3822]***
adj. R^2	0.57	0.78	0.76	0,78	0,78	0.78
F test	28.8391***	70.2724***	69.069***	61.53***	66.56***	61.1951***
SER	491,500	0.8007	0.8061	0.7682		0.7699
AIC	6634.69	557.385	560.473	541.3	539.358	542.32
SBC	6675.84	598.537	601.625	592.74	587.37	593.76
HQC	6651.29	573.989	577.076	562.054	558.73	563.075
RESET	0.5836	7.2237***	8.6588***	0.5302	0.1525	0.3196
RESET^{R}		9.3072***	10.1092***	0.6017	0.1059	0.2724

Standard error in () and standard error robust to het. in []. ^{T} LP and LTRB (transformed) and AGE not transformed

Accordingly, the mixed semilog with quadratic but no iteration term between the numerical independent variables is the candidate for a more exhaustive analysis. Notice that the price elasticity for TRB and semielasticity for AGE are not constant and that there is no interaction among them.

Once the mixed-semilog (4^R) has been selected, we wonder if the transformation we have a priori chosen for the dependent numerical variables (i.e the price and GT in logs and AGE untransformed) has any statistical support, or in other words, if there exist and alternative transformation pattern with a superior fit to the data. Our decisions will be based on the results of the *t* test of Davidson-MacKinnon (DM) [24] and F test of Mizon y Richard (MR) [25] and the, test that are appropriate to discriminate among any two non-nested models provided that they have the same dependent variable (in our case log(P) and/or P). In order to show the complete picture and compare alternative model configurations following the structure (4^R) with alternative transformation of the dependent variable, we are also including the results of the Box-Cox (1964) (BC) [26] and Wooldridge tests (W) [22]. Tables 4(a) and 4(b) summarise the results.

For one side, both MR[4] and DK[5] are coincident when discriminating among alternative models derived from the inclusion of numerical independent variables in log or in level. Thus, based on the test carried on in order to discriminate among the alternative non-nested models included in Table 4(a), a priori accepting the log(P) transformation for the dependent variable, (4^R2) is the one that fits the data best. For another side, BC[6] and W[7]

---

4 The M&R test works as follows. For example, in order to discriminate between Model 2 and Model 4 M&R test is based on the joint significance tests for the coefficients (AGE), (AGE)2, log(AGE) and log(AGE)2 derived from the estimation of the joint model including both, [AGE, AGE2] and [log(AGE) y log(AGE)2]. The joint test for AGE y AGE2 is significant [F(2, 212) = 3.34649 p-value 0.0370773], which implies that MODEL 2 is preferred. However, when discriminating between Model 4 and Model 2, the joint significance test of the coefficients for log(AGE) y log(AGE)2 [F(2, 212) = 2.88147, (p-value = 0.0582)], although significant, it's not as conclusive as the previous (5,8%).

5 The D&M test works as follows. For example in order to discriminate between Model 2 and Model 4 MR test is based on the individual t statistics derived from regressing Models 2 including also the fitted values coming from MODEL 4 as an additional independent variable. Since t is not significant for this added variable [t=1.6418 , p=0.1012], Model 2 offers a best fit. When following the same procedure, one discriminates between Model 4 and Model 2, the resulting t statistics is significant [t=2.4591, p=0.0147], which constitutes a clear evidence against Model 4.

6 We are using the version of the Box-Cox test (1964) developed by Zarembka (1968) to compare the linear and log transformations of the dependent variable. The underlying idea behind the procedure is to scale the observations on dependent variable (P_i) so that the residual sum of squares in the linear and logarithmic models are rendered directly comparable. The next steps are followed in order to calculate the test statistics reported in Table 4(b): 1) Scale the observations on P_i by dividing them by the sample geometric mean of P_i ($P_i^*=P_i/GMP_i$). 2) Regress the lineal model using P_i^* instead of P_i (Model 7) and the logarithmic model using

tests carried on to compare (4^R2) and the alternative one including the same right side but the untransformed P as dependent variable (4^R7) are also coincident. The W rejects the model including the untransformed vessel price (P) as dependent variables (4^R7), and gives statistical support to accept (more precisely not to reject) the model including Log(P) (4^R2); For another side, the Box-Cox test discriminates clearly in favour of the logarithmic specification (4^R2).

Table 4(a). Testing between the log vs no-transformation of the independent variables with Log(P) as dependent.

Models	Model	F MR	Decision	*t* DM	Decision
4^R2 vs 4^R4	4^R2	2.88147*	$4^R2 \succ 4^R4$	1.6418	$4^R2 \succ 4^R4$
	4^R4	3.34649**	$4^R4 \prec 4^R2$	2.4591**	$4^R4 \prec 4^R2$
4^R2 vs 4^R4 HetR	4^R2	5.06363***	$4^R2 \succ 4^R4$	1.557	$4^R2 \succ 4^R4$
	4^R4	6.35335***	$4^R4 \prec 4^R2$	3.010***	$4^R4 \prec 4^R2$
4^R4 vs 4^R5	4^R4	1.6957	$4^R4 \succ 4^R5$	-0.187	$4^R4 \succ 4^R5$
	4^R5	55.57***	$4^R5 \prec 4^R4$	14.401***	$4^R5 \prec 4^R4$
4^R4 vs 4^R5 HetR	4^R4	0.083789	$4^R4 \succ 4^R5$	-0.333	$4^R4 \succ 4^R5$
	4^R5	151.928***	$4^R5 \prec 4^R4$	17.438***	$4^R5 \prec 4^R4$
4^R2 vs 4^R5	4^R2	0.0713	$4^R2 \succ 4^R5$	-0.189	$4^R2 \succ 4^R5$
	4^R5	103.28***	$4^R5 \prec 4^R2$	14.397***	$4^R5 \prec 4^R2$
4^R2 vs 4^R5 HetR	4^R2	0.083789	$4^R2 \succ 4^R5$	0.404	$4^R2 \succ 4^R5$
	4^R5	151.928***	$4^R5 \prec 4^R2$	17.674***	$4^R5 \prec 4^R2$
4^R4 vs 4^R6	4^R4	0.052906	$4^R4 \succ 4^R6$	-0.165	$4^R4 \succ 4^R6$
	4^R6	99.61***	$4^R6 \prec 4^R4$	14.141***	$4^R6 \prec 4^R4$
4^R4 vs 4^R6 HetR	4^R4	0.0642879	$4^R4 \succ 4^R6$	-0.302	$4^R4 \succ 4^R6$
	4^R6	143.846***	$4^R6 \prec 4^R4$	17.055***	$4^R6 \prec 4^R4$
4^R2 vs 4^R6	4^R2	1.4536	$4^R2 \succ 4^R6$	t=0.366	$4^R2 \succ 4^R6$
	4^R6	52.5519***	$4^R6 \prec 4^R2$	14.397***	$4^R6 \prec 4^R2$
4^R2 vs 4^R6 HetR	4^R2	2.5456**	$4^R2 \succ 4^R6$	t=0.512	$4^R2 \succ 4^R6$
	4^R6	77.6172***	$4^R6 \prec 4^R2$	17.226***	$4^R6 \prec 4^R2$

Model 4^R2: $\log(P)=\beta_0+ \beta_1\log(GT)+ \beta_2\log(GT)^2+ \beta_3AGE+ \beta_4AGE^2+$etc.
Model 4^R4: $\log(P)=\beta_0+ \beta_1\log(GT)+ \beta_2\log(GT)^2+ \beta_3\log(AGE)+ \beta_4 \log(AGE^2)+$etc.
Model 4^R5: $\log(P)=\beta_0+ \beta_1(GT)+ \beta_2(GT)^2+ \beta_3(AGE)+ \beta_4(AGE^2)+$etc.
Model 4^R6: $\log(P)=\beta_0+ \beta_1(GT)+ \beta_2(GT)^2+ \beta_3\log(AGE)+ \beta_4\log(AGE^2)+$etc.

$\log(P_i^*)$ instead of $\log(P_i)$ (Model 2), but otherwise leaving the models unchanged. The RSS are now comparable, the model with the lower RSS (Model 2) providing the better fit. 3) To check if (Model 2) is providing a significantly better fit than the linear the $\chi^2(1)=$ (n/2)log(RSSlog/RSSlinear)=231, where n=228 is the number of observations in the sample. Under the null hypothesis that there is not difference, this statistics is distributed as a $\chi^2(1)$, and accordingly the Model 2 is provides a significantly better fit that Model 7.

[7] The procedure in Wooldridge (1994) can be summarised as follows: 1) Obtain the fitted values from the primary regression (PR) $\varphi(P_i)$ on the independent variables $(X_i) \rightarrow \hat{j}$. 2) Obtain the fitted values and the residuals from the inverse regression (IR) P_i on $\hat{j}^{-1}(\hat{j}_i) \rightarrow \hat{j}$, $\hat{e}_i$ and calculate the weighted residuals $\tilde{e}_i = \hat{e}_i / \hat{P}_i^{2(1-\lambda)}$, where λ=1 for the linear model and λ=0 for the log model. 3) For the linear model (λ=1) obtain the scalar residuals $\tilde{r}_i (= \hat{e}_i)$ from $\hat{P}_i \log(\hat{P}_i)$ on X_i; For the log model (λ=0) obtain the scalar residuals from $[\log(\hat{P}_i)]^2$ on X_i. 4) Compute the sum of square residuals (SSR) from the regression 1 on $\tilde{e}_i \cdot \tilde{r}_i$. N-SSR is distributed asymptotically as $\chi^2(1)$.

Table 4(b). Testing between Log(P) vs P with the right side as 4^R.

Models	Model	χ^2 Wooldridge	Decision	χ^2 Box-Cox	Decision
4^R2 vs 4^R7	4^R2	0.0762	Accept 4^R2	231.13***	Accept 4^R2
	4^R7	30.42***	Reject 4^R7		Reject 4^R7

Model 4^R2: $\log(P)=\beta_0+\beta_1\log(GT)+\beta_2\log(GT)^2+\beta_3AGE+\beta_4AGE^2+etc.$
Model 4^R7: $P=\beta_0+\beta_1(GT)+\beta_2(GT)^2+\beta_3\log(AGE)+\beta_4\log(AGE^2)+etc.$

3.2 Structural Change

Based on the analysis undertaken in the previous subsection the selected functional form (4^R2) is a mixed semilog model that includes quadratic but no interaction term between the numerical independent variables (GT and AGE):

$$\log(P_{it})=\beta_0+\beta_1\log(GT_{it})-\beta_2(AGE_{it})-\beta_3\log(GT_{it})^2+\beta_4(AGE_{it})^2+\beta_5(MAGP_{1A})+\beta_6(MAGP_{2B})+\beta_7(MAGP_{2A})+\beta_8(MAGP_{3B})+\beta_9(MAGP_{4B})+\beta_{10}(MAGP_{4A})+\beta_{11}(SUB_B)+\beta_5(SUB_A)+u_1 \qquad (4^R2)$$

where P_{it} represents the transaction price, GT_{it} are the gross tonnes of the vessel (a usual measure of the fishing capacity of a vessel), AGE_{it} is the years since the vessel was constructed when the transaction happens and, GT_{it}^2 and AGE_{it}^2 are their related quadratic terms. The rest of independent variables are derived from crossing the two dummies representing the SECTOR the purchased/sold vessel belongs to (artisanal, trawlers) and the MAGP under which the transaction took place ($MAGP_1$, $MAGP_2$, $MAGP_3$, $MAGP_4$, SUB). This way, vessels are classified in 10 sub-groups attending to SECTOR and TIME, being the artisan vessels sold during the $MAGP_1$ (i.e. $MAGP_{1B}$) the base case.

Table 5. Testing for Structural Change.

test*	Model	H_0	F & F^{R*}
1	4^R2'EXT	$\delta_1=0, \delta_2=0, \delta_3=0, \delta_4=0$	F(4,198)=0.2799 F^R(4, 198)=0.2393
2	4^R2'EXT	$\delta_9=0, \delta_{10}=0, \delta_{11}=0, \delta_{12}=0, \delta_{13}=0, \delta_{14}=0, \delta_{15}=0, \delta_{16}=0, \delta_{17}=0, \delta_{18}=0, \delta_{19}=0, \delta_{20}=0, \delta_{21}=0, \delta_{22}=0, \delta_{23}=0, \delta_{24}=0$	F(16,198)= 0.7430 F^R (16,198)=1.1114
3	4^R2'EXT	$\delta_1=0, \delta_2=0, \delta_3=0, \delta_4=0, \delta_9=0, \delta_{10}=0, \delta_{11}=0, \delta_{12}=0, \delta_{13}=0, \delta_{14}=0, \delta_{15}=0, \delta_{16}=0, \delta_{17}=0, \delta_{18}=0, \delta_{19}=0, \delta_{20}=0, \delta_{21}=0, \delta_{22}=0, \delta_{23}=0, \delta_{24}=0$	F(20,198)= 0.6603 F^R (20,198)=1.176
4	4^R2''EXT	$\delta_0=0, \delta_2=0, \delta_4=0, \delta_6=0, \delta_8=0, \delta_{13}=0, \delta_{14}=0, \delta_{15}=0, \delta_{16}=0$	F(9,210)=0.8239 F^R (9,210)=0.9291

F^{R*} = F robust to heteroskedasticity
Test 1: Is the effect of and additional GT and/or AGE on P the same for artisanal and trawling vessels?
Test 2: Is the effect of GT and AGE on P estable during the different MAGPs?
Test 3: Does the slope related to GT and/or AGE depend on both, sector and period?
Test 4: Are the estopes related to GT and/or AGE different for artisan and trawlers?

Although model (4^R2) contemplates structural change via changing constant term depending on SECTOR and MAGP, this subsection checks the convenience of including additional structural flexibility. In this sense, we are now interested on testing whether not only the constant term, but also the slopes of the estimated price function with respect to the numeric variables (i.e. GT and/or AGE) differ depending on $SECTOR_i$ and $MAGP_i$. If the answer was yes we should consider dividing the sample by sector and/or period and, accordingly estimate different price functions for each of the sub-samples. Attending to the resulting degrees of freedom, while dividing the sample by sector would be statistically tractable, however, fragmenting it by period would imply an insufficient sample size.

Even if (4^R2) is suitable to give answer to the questions raised in the beginning of the paper (that is to check if the European fisheries policy has influenced on the second market of fishing vessels and if this influence on prices depends on sub-sector) however is not the best choice to face the objectives of this subsection. Accordingly, we opt for an additional model (4^R2'), which, being completely equivalent to (4^R2), is however more appropriate to analyse if additional structural flexibility should be introduced. Model (4^R2') includes the numerical variables and their quadratic terms, the two dummies for $SECTOR_i$ and $MAGP_i$ and the interaction term between the dummies, that is $SECTOR_i * MAGP_i$.

$$\log(P)= \beta_0+\beta_1\log(GT)+\beta_2\log(GT)^2+\beta_3AGE+\beta_4AGE^2+\beta_5MAGP_2+ \beta_6MAGP_3+\beta_7MAGP_4+\beta_8SUB+\beta_9SECTOR+ \beta_{10}MAGP_2*SECTOR+ \beta_{11}MAGP_3*SECTOR+ \beta_{12}MAGP_4*SECTOR+\beta_{13}SUB*SECTOR+u \quad (4^R2')$$

$$\log(P)=\beta_0+\delta_0SECTOR+\beta_1\log(GT)+\delta_1\log(GT)*SECTOR+\beta_2\log(GT)^2+ \delta_2\log(GT)*SECTOR+\beta_3AGE+\delta_3AGE*SECTOR+\beta_4AGE^2+ \delta_4AGE^2*SECTOR+\delta_5MAGP_2+\delta_6MAGP_3 +\delta_7MAGP_4+\delta_8SUB+ \delta_9MAGP_2*LGT+ \delta_{10}MAGP_2*LGT2+ \delta_{11}MAGP_2*AGE+ \delta_{12}MAGP_2*AGE2+\delta_{13}MAGP_3*LGT+\delta_{14}MAGP_3*LGT2+ \delta_{15} MAGP_3*AGE+\delta_{16}MAGP_3*AGE2+\delta_{17}MAGP_4*LTRB+ \delta_{18}MAGP_4*L TRB2+\delta_{19}MAGP_4*AGE+\delta_{20}MAGP_4*AGE^2+ \delta_{21}SUB*LGT+ \delta_{22}SUB*LGT^2+ \delta_{23}SUB*AGE + \delta_{24}SUB*AGE^2+u \quad (4^R2'EXT)$$

In order to achieve the objective above mentioned we first estimate an extended version of model $4^R2'$, that is model ($4^R2'$EXT). In this extended model not only the constant terms (such as in 4^R2 and $4^R2'$) but also all the slopes are allowed to change depending on subgroups. In other words, model $4^R2'$EXT permits testing whether the price elasticity for GT (ε_{GT}) or the price semi-elasticity of AGE (ε_{AGE}) are equal for artisan and trawling vessels (Test 1) or whether ε_{GT} or ε_{AGE} are stable (i.e. if the slope of the price function changes with time (Test 2), as wells as if the slope of the price function depend both on SECTOR and MAGP (Test 3). The F statistics for each of the tests (Table 5) are non significant. This means that the interaction terms between the numerical and dummies are irrelevant to explain prices, or to put in another words, that the slope of price respect GT and AGE are the same for inshore and offshore vessels and that they remain constant during the different MAGPs periods.

Once we have rejected the extended version of the models 4^R2 and $4^R2'$, for completeness, we are also checking if there are structural differences between artisan and trawling vessels, that is, if the slopes for the independent variables (ie. GT and AGE) are different by sector (test 4). Based on an additional model ($4^R2''$EXT) the F test is performed. Since the resulting F statistics is not significant there is not evidence against working with the pooled sample. This allows working with more degrees of freedom.

$$\log(P)=\beta_{0+}\delta_0 SECTOR+\delta_1 LGT+\delta_2 LGT*SECTOR+\delta_3 LGT^2+\delta_4 LGT^2*SECTOR+\delta_5 AGE+\delta_6 AGE*SECTOR+\delta_7 AGE^2+\delta_8 AGE^2*SECTOR+\delta_9 MAGP_2+\delta_{10} MAGP_3+\delta_{11} MAGP_4+\delta_{12} SUB+\delta_{13} MAGP_2*SECTOR+\delta_{14} MAGP_3C*SECTOR+\delta_{15} MAGP_4C*SECTOR+\delta_{16} SUB*SECTOR \quad (4^R2''EXT)$$

3.3 Alternative Estimation Methods

The specification of the functional form of hedonic regression models, the configuration of models suitable to test for the structural break, and the discussion of the appropriate estimation method (often OLS *vs* WLS) have been the subject of considerable attention in the hedonic literature. This subsections deals with the third issue by presenting the results derived from using three alternative estimations methods to estimate the mixed semi-logarithmic model selected in the previous subsections (4^R2). These methods are: ordinary least squares (OLS), weighted least squares (WLS) and trimmed least squares (TLS).

Table 6. Font sizes of headings. Table captions should always be positioned *above* the tables.

Test	Potential Origin	Statistics	P-value
t test	LGT	t=-0.5215	(0.6025)
t test	LGT^2	t=-0.7432	(0.4581)
t test	AGE	t=1.590	(0.1132)
t test	AGE^2	t=1.274	(0.2041)
t test	$MAGP_{1A}$	t=-0.6970	(0.4865)
t test	$MAGP_{2B}$	t=2.867	(0.0045)***
t test	$MAGP_{2A}$	t=1.438	(0.1519)
t test	$MAGP_{3B}$	t=0.7320	(0.4649)
t test	$MAGP_{3A}$	t=-0.7111	(0.4778)
t test	$MAGP_{4B}$	t=-2.084	(0.0382)**
t test	$MAGP_{4A}$	t=-0.9452	(0.3456)
t test	SUB_B	t=-0.7335	(0.4640)
t test	SUB_A	t=0.3266	(0.7442)
F tets	$MAGP_{2B}$ $MAGP_{4B}$	F(2, 225)=5.3900	(0.0051)**
F test	LGT, LGT^2	F(2, 225)=0.8599	(0.4245)
F test	AGE AGE^2	F(2, 225)=1.8508	(0.1594)
F tets	GT, LGT^2 AGE AGE^2	F(4, 223)=0.4644	(0.7616)
F test	$MAGP_{ij}$ (i=1…5) (j=1,2)	F(9, 218) = 1.7645	(0.0763)*
F tets	LGT, LGT^2 AGE AGE^2 $MAGP_{ij}$	F(13, 214) = 1.8883	(0.0327)**
LM Koenker (BP)	LGT, LGT^2 AGE AGE^2 $MAGP_{ij}$	LM(χ^2(13))=23.4629	(0.0364)**
White	LGT, LGT^2 AGE AGE^2 $MAGP_{ij}$	χ^2(57)=69.6047	(0.1220)
Reduced White	$\hat{y}$, $\hat{y}^2$	F(2, 225) = 1.6039	(0.2233)

In order to judge the convenience of using WLS instead of OLS to remove heteroskedasticity of variances, alternative test have been carried on to detect it. Table 6 summarises the results including also the respective potential origin. For one side, both the F and Breusch-Pagan test are significant, which implies that het might bias the inference based on OLS standard errors. However, for another, White and reduced White tests do not detect it. Thus, since the results of the tests are inconclusive, we are also reporting the WLS estimates. The approach summarised in Woldridge (2003, pag. 306) has been followed to calculate the weighted factors.

Last but not least, the regression diagnostic has been carried out in order to see if there are atypical or/and influential observations in the data set with a powerful influence on the estimated parameters and/or predictions of the model. Once selected these will be candidates to be omitted to deal with robust regression techniques. Usual measures in regression diagnostic (i.e. the leverage (ht), Cook's distance (CD), the standardised residual (e*), DFBETAs and DFFITS are reported in Table 7. All the data points exceeding any of the rule-of-thumb cut-offs of the above-mentioned measures have been market with an asterisk. Six data points (market with double asterisk (0bs coded: 12, 112, 145, 244, 246 and 294)) either with a moderate leverage and/or outliers or a significant contribution on the values of the estimated parameters and/or the model's predictions have been detected, which represents the 2.63% of the sample data[8]. The trimming proportion that guaranties the elimination of all

[8] While not necessarily undesirable, influential observations are those observations that make a relatively large contribution to the values of the estimates, that is, observations whose inclusion or exclusion may result in substantial changes in the fitted model. The most common measures for the degree of influence are the *leverage* (ht) and to some degree *Cook's distance* (CD). As a general rule, data points satisfying [0.2>ht>0.5] are considered moderately influent, while those in which ht>0,5 should be especially kept watch. The sample size corrected rule of thumb suggested by Belsey et al. (1980) is h>2p/N, where p is the number of estimated parameters and N the sample size. Similarly, the general criterion stands to watch out for observations where CD>1, although in large samples some authors suggest a sample corrected rule of CD>4/N. Applying these rules to our case study, hardly 2.19% of the observations are moderately influent attending to ht, while the 8.3% does go beyond the sample corrected rule relative to CD. Thus, it may be concluded that none of the observations are riskily influent according nor to the leverage, nor the CD. Together with influential observations, it is also convenient to include measures designed to detect large errors. In a model which fits in every cell formed by the independent variables, no absolute standardized residual will be e*>2 (0,05 level) (or e*> 1.96(0.01 level)). Cells not meeting this criterion indicate combinations of independent variables for which the model is not working well. In our model about 94 % of the observations fit the specified rule of thumb, and consequently regarded as acceptable in term of the model specification. Thus, about 6 % of the observations may be considered outliers. Outliers and high leverage points can be an indication of exceptional data points that are worthy of further study. What is likely to be of more importance however is whether these points significantly contribute to the values of the

the influential outliers and high leverage observation is 0.05, which implies the consideration of the residuals associated with the 0.05 and 0.95 quartiles. In addition to all the influential outliers and high leverage observations, the TLS procedure has picked up some others of moderate size. Therefore, those observations where the residuals are non-positive for ?=0.05 and non-negative for ?=0.95 have been discarded. Subsequently, least squares have been applied to the remaining observations.

Table 7. OLS Regression Diagnostic.

Obs.	e.stand.	e.stud.	CD	h	DFFITS	DFB_6	DFB_7	DFB_8	DFB_{10}	DFB_{12}	DFB_{13}
1	-1,6350	-1,9697	0,0248*	0,0721	-0,5945*	0,1095	0,0895	0,1117	0,11319	0,1066	0,0892
12**	-1,3683	-1,6456	0,0177*	0,0736	-0,5010*	0,0937	0,1116	0,0893	0,09149	0,0872	0,1138
15	2,0696*	2,5028	0,0380*	0,0686	0,7395*	-0,132*	-0,1227	-0,135*	-0,134*	-0,138*	-0,117
46	-2,0289*	-2,5196	0,0240*	0,0436	-0,5881*	-0,0837	0,0159	-0,0027	0,0000	-0,0036	0,0172
50	-3,1957*	-3,9196*	0,0742*	0,0554	-1,0584*	-0,1289	-0,0535	0,01517	0,0181	0,0110	-0,0420
52	-1,6789*	-2,0806	0,0171	0,0456	-0,4941	-0,0681	0,0184	-0,0015	0,0008	-0,0016	0,0215
57	0,5777	0,6499	0,0055	0,1332*	0,2792	0,0238	0,0024	-0,0020	0,0012	-0,0049	0,0022
83	-2,665*	-3,2896*	0,0467*	0,0497	-0,8294*	-0,1103	-0,0407	0,00766	0,0105	0,0029	-0,0368
86	2,1231*	2,6289	0,0278*	0,0464	0,6339*	0,0870	-0,0024	-0,0022	-0,0054	-0,0046	-0,0092
109	-4,1726*	-5,2766*	0,0648*	0,0262	-1,0214*	0,0005	-0,0355	-0,0907	0,0094	-0,0062	-0,0350
112**	-1,7315*	-2,0736	0,0298*	0,0775	-0,6523*	-0,0090	-0,0276	-0,0391	-0,0073	0,0160	-0,0331
129	1,0793	1,3358	0,0072	0,0468	0,3194	-0,0035	-0,0003	0,02235	-0,0011	0,0039	-0,0001
136	1,1319	1,3937	0,0087	0,0517	0,3507	-0,0049	-0,0009	0,0225	-0,0029	0,0046	-0,0021
145**	-1,6473*	-2,0312	0,0180*	0,0504	-0,5072*	0,0034	0,0177	-0,0339	0,0057	0,0100	0,0292
217	1,0577	1,05644	0,0318*	0,2299*	0,6687*	-0,0028	0,0259	-0,0135	-0,0131	0,0261	0,0098
222	-1,9126*	-2,3132	0,0324*	0,0685	-0,6814*	0,0056	0,0439	-0,0043	-0,0005	-0,1017	0,0468
244**	-2,0432*	-2,4039	0,0498*	0,0937	-0,8458*	-0,0014	-0,209*	0,00199	0,0034	-0,0008	-0,0250
246**	-2,0529*	-2,4166	0,0500*	0,0932	-0,8477*	-0,0002	-0,2081*	0,00289	0,0048	-0,0002	-0,020
250	1,3143	1,5544	0,0196*	0,0890	0,5260*	-0,0017	0,1134	-0,0001	-0,0010	0,0023	-0,0071
251	-0,8134	-0,8915	0,0128	0,1555*	-0,4242	-0,0032	-0,071	-0,0012	-0,0047	0,0077	0,0018
265	0,2249	0,2539	0,0008	0,1304*	0,1075	-0,0004	-0,0094	0,00104	0,0017	0,0002	-0,0092
271	-2,3670*	-2,9650	0,0270*	0,0354	-0,6280*	0,0018	0,0021	0,00057	0,0020	-0,0030	0,0033
289	-0,2830	-0,3207	0,0012	0,1270*	-0,1336	0,0011	0,0096	-0,0007	-0,0008	-0,0014	0,0100
294**	-2,1198*	-2,5071	0,0510*	0,0890	-0,8572*	0,0015	-0,0077	0,00327	0,0047	0,0023	-0,1966*
298	1,3083	1,53689	0,0207*	0,0951	0,5405*	0,0029	0,0121	0,00168	0,0020	0,0025	0,1357*
299	1,3083	1,53689	0,0207*	0,0951	0,5405*	0,0029	0,0121	0,00168	0,0020	0,0025	0,1357*

The estimations of the mixed semi-log (i.e model 4^R2) using the three alternative methods are reported in Table 8. The comparison of OLS, WLS and TLS estimators allows one to conclude that the differences are not highly significant. This in turn may demonstrate that the estimated parameters by

coefficient estimates and the model predictions. Diagnostics respectively designed for these two purposes are DFBETAS and DFFITS (Belsey et al., 1980). The general cut-off criterion for cases to be considered forceful to the values of the coefficients is $|DFBETAS_{kt}| > 1.0$ (Menar, 1995), while Belsey et al. recommend further investigation of observations where $|DFBETAS_{kt}| > 2/N^{(1/2)}$, specially in big samples. Regarding the predictions, the general rule stands that an observation is considered forceful to the predictions when $|DFFITS_{kt}| > 1$, while the sample corrected rule is $DFFITS_{kt}| > 2/(p/N)^{(1/2)}$. All the data points exceeding the rule of thumbs have been market with an asterisk.

OLS are robust to using additional procedures. Nevertheless the elasticity estimation and price patterns using the alternative estimated functions will be also compared.

Table 8. Regression Results Using OLS, WLS and TLS.

Variable	OLS	WLS	TLS*
constant	6.7621 (0.4204)***	6.63155 (0.3713)***	6.3885 (0.3265)***
LGT	2.0560 (0.2203)***	2.13523 (0.1634)***	2.2181 (0.1714)***
AGE	-0.0813 (0.0268)***	-0.0790 (0.0166)***	-0.046780 (0.0049)**
LGT^2	-0.1507 (0.0268)***	-0.1621 (0.0233)***	-.18333 (0.0208)***
AGE^2	0.0011 (0.00067)*	0.0012132 (0.0004)****	0.0049039 (0.0052)
$MAGP_{1A}$	0.1058 (0.3500)	0.116813 (0.3425)	.22671 (0.2715)
$MAGP_{2B}$	0.6272 (0.2536)**	0.796769 (0.3290)**	.66758 (0.1964)***
$MAGP_{2A}$	0.6094 (0.3314)*	0.364523 (0.5856)	1.10247 (0.2571)***
$MAGP_{3B}$	1.1668 0.2311***	1.04569 (0.2550)***	1.1595 (0.1793)***
$MAGP_{3A}$	1.6182 (0.2998)***	1.65537 80.2795)***	1.8134 (0.2326)***
$MAGP_{4B}$	1.3071 (0.2261)***	1.2601 (80.2410)***	1.1481 (0.1754)***
$MAGP_{4A}$	1.7253 (0.3368)***	1.6832 (0.2820)***	1.7903 (0.2612)***
SUB_B	1.3046 (0.2627)***	1.14844 (0.2522)***	1.3037 (0.2038)***
SUB_A	1.9834 (0.3352)***	1.98734 (0.3631)***	1.9934 (0.2600)***
adj. R^2	0.78	0.85	0.79

*Number of observations after trimming= 189

4 Interpreting the Results

Based on the analysis carried out in section 3, the OLS estimations for the mixed semilog function (4^R2) will be used to answer the researching questions raised in the beginning. Mainly, if the European Fisheries Structural Policy

has influenced on the second hand market price of Basque fishing vessels and/or if the influence differs according to the fishing subsector (i.e. artisanal, trawlers).

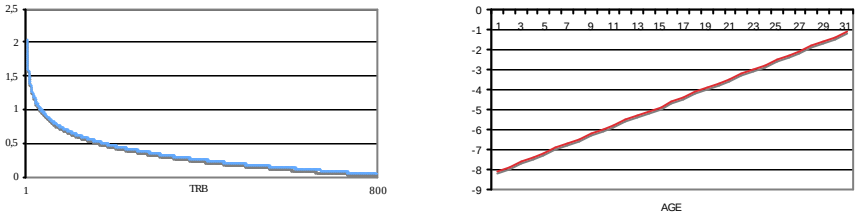
One of the main advantages of using a quadratic functional form such as (4^R2), is that since it doesn't assume constant elasticities and semi-elasticities for GT (ε_{GT}) (7) and AGE (s-ε_{AGE}) (8), it allows capturing increasing or decreasing marginal effects on vessel prices. When introducing the quadratic term for GT (β_3) and AGE (β_4) we are in fact allowing their respective price elasticity and semi-elasticity to change with GT and AGE values. Notice also that, since the TRANSLOG and semi-TRANLOG structures have been rejected, there is no interaction term between GT and AGE. Accordingly, ε_{GT} and s-ε_{AGE} are independent of the values of each other. To put in other words, ε_{GT} only depends on the values of GT and ε_{AGE} on its own values.

$$\varepsilon_{GT} = [\beta_1 - 2\beta_3 \log(GT)] = [2{,}02 - 0{,}28 * \log(GT)] \quad (7)$$

$$\varepsilon_{AGE} = \%\Delta price/\Delta AGE \approx 100 * \{[\beta_2 + 2\beta_4]AGE\} = (-8{,}13 + 0{,}23 * AGE) \quad (8)$$

Figure 2 illustrates the ε_{GT} relative to the range for GT values for the Basque fleet. Taking into account that the range of GT for the artisan and trawlers is respectively between GT_B=[2.15, 153] and GT_A=[240-793], the estimated ranges for the price elasticity with respect GT are: ε_{GTB}(B)=[2,15-0,53] for the artisan vessels and ε_{GTA}(A)=[0,40-0,04] for the trawlers.

Figure 2. Price elasticities and semielasticities related to GT (eTRB) and AGE (eAGE).



ε_{TRB} ε_{AGE}

Since s-ε_{AE} is a semi-elasticity, notice that it represents the approximated perceptual change on the vessel price as a result of an additional unit in AGE. Just like for GT, when including a quadratic term for AGE, the perceptual chance on the price of a year older second hand vessel, instead of being constant depends on the values of AGE. Notice that since the coefficient for AGE (β_2) is negative while the coefficient for the quadratic term (AGE^2) (β_4)

is positive, it may indicate that for small AGE values (almost new vessels) the fact that the vessel has an additional year has a negative effect on log(price). For AGE exceeding a critical value ($\beta_2/2\beta_4$ = 0,00004), this effect turns into positive. The shape of the quadratic form related to AGE means that the semielasticity of price with respect to AGE is increasing as AGE does. Thus, for example an increase of AGE from 5 to 6 years would decrease the price of a second hand market vessel in 6.73%. Taking into account that the sample average AGE is AGE=18, a one year older vessel would reach a price 3.69 % lower. Accordingly, it seems that the perceptual chance on prices as a result of a perceptual chance in AGE is sensibly higher for the newer vessels.

As well as analyzing the performance of the numerical variables (GT and AGE), the estimated model allows capturing the differences on prices related to the ten categories resulting from crossing the two dummy variables (i.e. $SECTOR_i$ and $MAGP_i$). This modeling approach, not only possibilities searching the price differences among the resulting ten categories, but also allows making significance test about the estimated differences in order to see if they are statistically significant or not.

For dealing with this, the subgroup of artisan vessels sold during $MAGP_1$ has been set up as the base category (i.e. $MAGP_{1B}$). Hence, the estimated coefficients for the rest of the $MAGP_{ij}$ measure the proportional difference on transaction price of a vessel sold during the $MAGP_{ij}$ (i?1, j?B) with respect to an artisan vessel sold during $MAGP_1$, keeping the same levels for GT and AGE. Table 9 includes all the estimated proportional differences among the categories and the results of the significance test carried on to determine if the differences are statistically significant.

The artisan vessels sold during $MAGP_2$ reached 62% higher prices than the ones sold during $MAGP_1$. The trawlers sold during the validity of $MAGP_2$ were approximately 10.5% more expensive than the artisan vessels during the same period. Likewise, the prices of the artisan vessels sold during the $MAGP_2$ were %53 higher than the ones sold during the previous period. At a first glance, the raising on price for the artisan vessels happened during $MAGP_2$ stands out, a period in which the trawlers were sold at 100% higher prices.

Table 9. Estimated Proportional Price Differences among Sub-groups.

Difference	Variation	Std. Error	t-statistic	p-value
$\Delta\log(p)[MAGP_{2B}\text{-}MAGP_{1B}]$	0,6272	0,2536	2,4733	0.0141**
$\Delta\log(p)[MAGP_{3B}\text{-}MAGP_{2B}]$	0,5395	0,1964	2,7468	0,0065***
$\Delta\log(p)[MAGP_{4B}\text{-}MAGP_{3B}]$	0,1403	0,1559	0,9003	0,369
$\Delta\log(p)[SUB_{B}\text{-}MAGP_{4B}]$	-0,0025	0,2020	0,0125	0,9900
$\Delta\log(p)[MAGP_{2A}\text{-}MAGP_{1A}]$	0,5036	0,3449	1,4602	0,1457
$\Delta\log(p)[MAGP_{3A}\text{-}MAGP_{2A}]$	1,0087	0,2794	3,6096	0,00038***
$\Delta\log(p)[MAGP_{4A}\text{-}MAGP_{3A}]$	0,1071	0,2701	0.3965	0,6921
$\Delta\log(p)[SUB_{A}\text{-}MAGP4_{A}]$	0,2581	0,3211	0,8037	0,4224
$\Delta\log(p)[MAGP_{1A}\text{-}MAGP_{1B}]$	0,1058	0,3500	0,3024	0,7626
$\Delta\log(p)[MAGP_{2A}\text{-}MAGP_{2B}]$	-0,0177	0,3055	-0,0580	0,9536
$\Delta\log(p)[MAGP_{3A}\text{-}MAGP_{3B}]$	0,4514	0,2614	1,7260	0,0857*
$\Delta\log(p)[MAGP_{4A}\text{-}MAGP_{1B}]$	0,4181	0,2935	1,4240	0,1557
$\Delta\log(p)[SUB_{A}\text{-}SUB_{B}]$	0,6788	0,3249	2,0890	0,0378**

The calculation of the standard errors for the difference on prices between groups (ij) needs further explanation. Notice that the estimated equation (4^R2) cannot be used to test for the statistical significance of such differences. The easiest way to execute this issue is by re-estimating the equation changing the base category in favour of one of the two categories whose differences one aims to check. Substantially nothing relevant changes, and this way, the required estimated values for the differences and the standard errors are directly obtained to conduct the related t tests. For example, focusing on the transactions of the artisan vessels taking place during $MAGP_2$, the *t* statistics to test the nil hypothesis H_0 that there is not difference on prices between the artisan vessels sold during $MAGP_{2B}$ and $MAGP_{1B}$ is t=0,6272/0,2536=2,47 (0,01417**), implying evidence against H_0.

It is worth pointing out that the differences on the second hand market of fishing vessels prices found to be statistically significant are $MAGP_2$ and $MAGP_3$ for the case of artisan vessels, and $MAGP_3$ for the trawlers. So there seems to be a link between the effective adjustment pattern happening in each sub-sector and the market prices of the second hand market vessels, because precisely, in the periods when the adjustments for each fleets have been most radical (i.e. MAGP2 and MAGP3 for the artisan fleet and MAGP3 for the trawlers) the price variations has been also higher, not only in magnitude but also in statistical significance. What all of this may be indicating? Since the fishing rights are concentrated in the hands of a lower number of hands, or to put in another words, since the remaining vessels increase their share in potential access rights, the resulting transaction prices my be force to increase.

For one side the shortage may induce price to move up. For another the gains in market power of the vessel owners may have also plaid a roll.

5 Conclusion and Economic Policy Recommendations

The result of this study suggests that there is a link between the evolution of the second hand market of fishing vessels and the European Fisheries Structural Policy. Effective capacity adjustments joint with a progressive tighten up of the requirements to access to European fishing grounds seems to have increased the hedonic price of a GT unit. Since building a new vessel requires the withdrawal of another with al least the same capacity, the second hand market of vessels stops being a mere market to buy-sell an asset and it in fact becomes in a market were fishing rights are exchanged. Evidence of this is that based on the applied hedonic model developed for the second hand Basque fishing vessels, the higher and statistically significant increases of prices precisely happens under the MAGP with major capacity adjustment: $MAGP_2$ and $MAGP_3$ in the case of inshore or artisan vessels and $MAGP_3$ in the case of offshore vessels (trawlers). This gives support to accept our hypothesis. Concretely our model deduces that: a) The inshore vessels sold during $MAGP_2$ reached 62%*** higher prices than the ones sold during $MAGP_1$. b) The inshore vessels sold during $MAGP_3$ reached a 53%*** higher price than the ones sold during $MAGP_2$. F) The offshore vessels sold during MAGP3 reached a 100%*** higher price than the ones sold during $MAGP_2$. K) The offshore vessels sold during $MAGP_3$ reached a 45%* higher price than the inshore ones sold during $MAGP_3$. M) The offshore vessels sold during SUB reached a 67%** higher price than the ones sold during $MAGP_1$. Thus taking into account that there is a narrow link between fisheries policy and the second hand market of vessels, policy makers should take into account the extra surplus that is being transferred to the vessels owners via second hand market when calculating the amount of decommissioning grant per gross tonnage. This way they may succeed in considerable budget savings.

References

1. Lancaster, K. (1966): Change and innovation in the technology of consumption. *American Economic Review* (1966).
2. Griliches, Z.: Hedonic price indexes for automobiles: an econometric analysis of quality change, Cambridge (MA). Harvard University Press (1971).
3. Rosen, S. (1974): Hedonic prices and implicit markets: production differentiation in pure competition. *Journal of Political Economy*. 82, 34-5 (1974).

4. Bayley, M.J., Muth, R.F. and Nourse, H.O.: A regression method for real estate price index construction. *Journal of the American Statistical Association*. 58(304), 933-942 (1963).
5. Boskin, M.J., Dulberger, E.R., Gordon, R.J., Griliches, Z. and Jorgenson, D.W.. Towards a more accurate measure of the cost of living. Final report to the Senate finance committee from the advisory Commission to study. The consumer price index, December 4). Senate Finance Committee. Washinton, DC (1996).
6. Van Dalen, J. and Bode, B.: Quality corrected price indexes: the case of Dutch new passenger car market, 1990-1999. Applied Economics. 36, 1169-1197 (2004).
7. Reis, H.J. and Silva, S.: Hedonic price indexes for new passenger cars in Portugal (1999-2001). Economic Modelling 23, 890-908 (2006).
8. Yu, K. and Prud'homme, M.: Econometric issues in hedonic price indixes: The case of Internet service providers. Paper presented at the Brookings Workshop on communications output and productivity (2007).
9. Williams, B. (2008): A hedonic model for Internet access service in the Consumer Price Index. Monthly Labour Review, July 2008, 33- 48 (2008).
10. Chow, G.: Technological Change and the Demand for Computers. *American Economic Review*. 57: 1117–30 (1967).
11. Berndt, E. R. and Neal J. R.: *Hedonics for Personal Computers: A Reexamination of Selected Econometric Issues*, presented at "R&D, Education and Productivity", an international conference in memory of Zvi Griliches. Paris, France (2003).
12. Leishman, 2001: House building and product differentiation: an hedonic price approach. Journal of Housing and the Built Environment. 16:131-152 (2001).
13. McCluskey, J.J. and Rausser, G.C.: Hazardous waste sites and housing appreciation rates. *Journal of Environmental Economics and Management*. 45, 166-176 (2003).
14. Bourassa, S.C., Hoesli, M. and Sun, J.: A simple alternative house price index method. *Journal of Housing Economics*. 15, 80-87 (2006).
15. Shimizu, C. and Nishimura, K.G.: Pricing structure in Tokyo metropolitan land markets and its structural changes: pre-bubble, bubble, and post-bubble periods. *Journal of Real Estate Financial Economics*. 36, 475-496 (2007).
16. Li, W., Prud'homme, M. and Yu, K.: Studies in hedonic resale housing price indexes. Paper presented to the Canadian Economic Association 40th Annual Meetings. Concordia University, Montréal (2006).
17. Triplet, J.: *Handbook on hedonic indexes and quality adjustments in price indexes*. OECD Publishing (2006).
18. Smith, B.A. and Tesarek, W.: House Prices and Regional Real Estate Cycles: Market Adjustments in Houston. *Real State Economics*. Volume 19 Issue 3, Pages 396 – 416, (2001).
19. Hidano, N.: The economic valuation of the environment and public policy: a hedonic approach. Northampton, MA: Edward Elgar (2003).
20. Kiel, K.A., McClain, K.T.: House prices during siting decision stages: the case of an incinerator from rumor through operation. *Journal of Environmental Economics and Management*. 28, 241-255 (1995).
21. Barzyk, F.: Updating the hedonic equations for the price of computers. Working paper of Statistics Canada. Prices Division. November.
22. Wooldridge, J.M. (1994): A simple specification test for the predictive ability of

transformations models. *Review of Economics and Statistics*. 76, 59-65.
23. Malpezzi, S.: Hedonic pricing models: a selective and applied review. In: O'Sullivan, T., Gibb, L. (Eds.). Housing Economy and Public Policy. Blackwell, Malder, MA (2003).
24. Davidson, R. and MacKinnon, J.G.: Several tests of model specification in the presence of alternative hypothesis. *Econometrica*. 49, 781-793 (1981).
25. Minzon, G.E. and Richard, J.F.: The encompassing principle and its application to testing non-nested hypotheses. *Econometrica*. 54, 657-678 (1986).
26. Box, E.P. and Cox, D.R.: An analysis of transformations. *Journal of the Statistics Society Series*. 26(2): 211-243 (1964).

Vertical integration in the fishing sector of the Basque Country: applications to the market of mackerel

Javier García Enríquez

Department of Applied Economics III (Econometrics and Statistics)
University of the Basque Country
Avda Lehendakari Aguirre, 83
48015 Bilbao (Spain)
e-mail: javier.garcia@ehu.es
tel: (34) 94 601 3773

Abstract. This paper analyses the internal market of fresh fish at the Basque Country, in particular the prices of mackerel coming from the inshore sector. The existence of vertical equilibrium relationships between the prices of mackerel sold at auction (at ports) and sold at central market is examined. For that purpose, different econometric techniques for the analysis of market integration are applied, paying especial attention to the long-term equilibrium relationships by means of both integer and fractional cointegration. The results show that it is a non-integrated market, and each price series, at port and at central market, has an independent evolution.

Keywords: Integrated markets, structural change, cointegration, mackerel.

This paper examines the possibility of vertical equilibrium relationships in an internal market of fresh fish at the Basque Country. In particular we analyse the monthly prices of mackerel, coming from the inshore sector, sold at auction (at ports) and at the central market, from January 1997 to December 2006. Different econometric techniques for the analysis of market integration are applied, examining the long-term equilibrium relationships by means of both integer and fractional cointegration.

A necessary condition for cointegration is that the potentially cointegrated series be balanced, i.e. they should have the same integration order. Also, as a previous step to that verification, it is convenient to make sure that the series does not have any structural change, which can mislead the conclusions in favour of the I(1) hypothesis. The series of prices in Fig 1 show an apparent change in the structure of prices at the central market, which goes from being quite steady before 2002 to have an increasing tendency from 2002 onwards.

To confirm this visual impression we use the test of structural change proposed by Zivot and Andrews [1], which finds evidence of a structural change in the trend in November/December, coinciding with the first effects of the euro entrance as the European single currency (January 2002) and the sinking of the oil ship Prestige by the Spanish coast (November 2002).

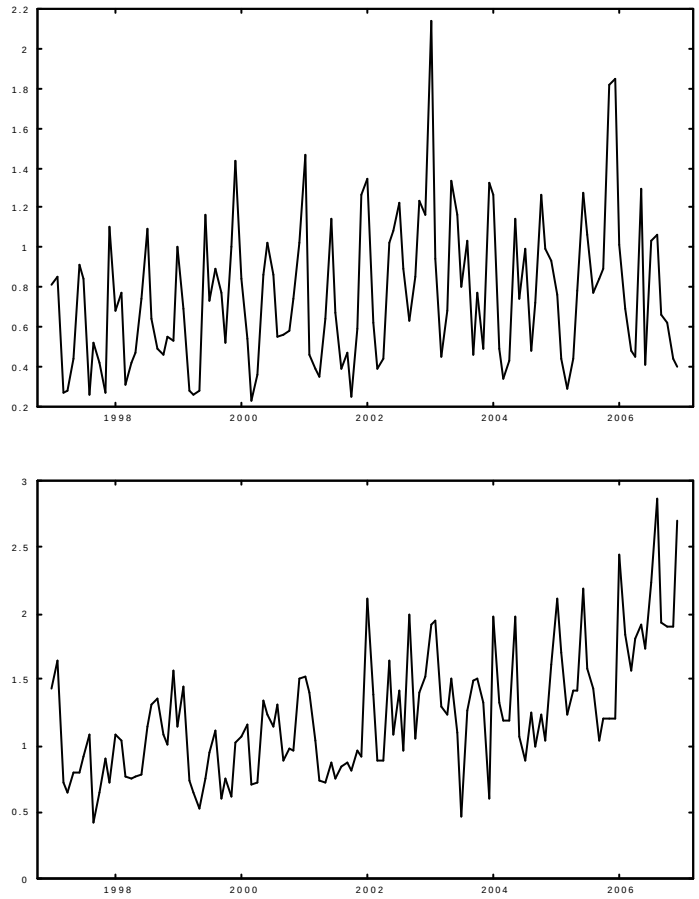


Fig 1. Prices of mackerel at ports (above) and at central market (below)

Once the structural change has been identified, it is removed by a simple regression in Gretl of the central market price series on a dummy trend, which takes zero until December 2002 and t, being t=1,2,…, from January 2003 until 2006.

Once the structural change has been removed, we implement the tests that will help us to determine the order of integration of the series. For that purpose, we use the Augmented Dickey and Fuller test [2] and [3] (hereafter ADF) and the

Kwiatkowski, Phillips, Schmidt and Shin stationarity test [4] (hereafter KPSS), both implemented in Gretl. Whereas the ADF test is not affected by the strong seasonality of the data, the results obtained with the KPSS test can be seriously distorted. For that reason, we first remove the seasonality in order to focus on the cointegration at frequency zero, which is the frequency related with a long run equilibrium.

Due to the limited number of observations, the estimation of the memory parameter at different cycles and the tests to detect the presence of seasonal unit roots are not very reliable. For that reason, we naively consider unit roots at all seasonal frequencies, and apply the summation filter over the corresponding series: $1+L+L^2+L^3+L^4+L^5+L^6+L^7+L^8+L^9+L^{10}+L^{11}$, where L is the lag operator. Notice that this filter is easily applicable in Gretl, cause it is only a sum of lagged series.

After the series have been seasonally adjusted, the ADF and KPSS tests are carried out. Both tests agree in not rejecting the I(1) hypothesis in the price series at ports at 5% significance level. However, the results for the price series at central market are more ambiguous. Whereas the ADF does not reject the unit root null hypothesis at 5% level, the KPSS test identifies the series as I(0) at 5% and I(1) at 10%. It is not possible, thus, to assure the existence of unit root on the basis of these tests. Therefore, in order to shed more light on the situation, we estimate the memory parameter of the differenced price series (to avoid nonstationarities) at central market by means of the semiparametric estimator of Geweke and Porter-Hudak [5] (hereafter GPH). A bandwdith $m=\lfloor T^{0.6} \rfloor$ is chosen, for T the sample size, because this is the value predetermined by Gretl. We obtain that $d=0$ falls in the 95% confidence interval, giving further evidence of unit root also in the price series at central market. In consequence, a cointegration analysis between these two series can be proposed.

In order to analyse the cointegration relationships, the two most usual tests in econometrics are used, that is, the Engle-Granger [6] and the Johansen [7] tests, both implemented in Gretl.

Taking into account that the direction of causality in the Engle-Granger test is unknown, we prefer to consider the two possible situations: the one in which the central market prices have an influence in the ports ones and vice versa. After applying the ADF test on the residuals of both cointegration relationships, evidence in favour of the I(1) hypothesis is confirmed.

Again, the results with the KPSS are more ambiguous. Whereas the stationarity of the cointegrating residuals is rejected at 5% significance level in the first causality specification, the results are less conclusive in the case of causality ports-central market, with a null hypothesis which is rejected at 10%

but not at 5%. Again we estimate the memory parameter of the first differences of the cointegrating residuals by means of the GPH method, and the value $d = 0$ falls inside the corresponding confidence interval, supporting the unit root hypothesis.

Thus, according to the Engle-Granger method, it does not exist neither fractional nor integer cointegration, since the residuals in both cointegration relationships can be considered integrated of the same order as the series, that is, I(1).

Finally, the Johansen's trace and maximum eigen value tests are applied. The hypothesis of zero cointegrating relationships is not rejected in any case, giving further evidence that the mackerel prices series at ports and at the central market are not cointegrated.

The methods to detect cointegration show that, even in the absence of structural change in the central market, the two studied links of the mackerel value chain in the Basque Country are not integrated, taking each one an independent evolution. The regulation systems of this fishery have an influence in the configuration of both markets (ports and central market), contributing to very different process in the fixing of prices. The fish prices at ports depend on aspects such as the Total Allowed Captures (hereafter TAC), the bilateral agreements with other countries, the fishing zones or areas, the fishing techniques used, etc. In the particular case of the mackerel, the auctioned amounts at ports do not depend as much on the demand as on the fixed TAC, and on the consecutive agreements, becoming a rigid supply. However, the prices at central market seem to answer a traditional demand and supply model. Thus, the result here described is not surprising. Whereas the central market has managed to transmit to the prices the shocks that have affected negatively to this fishery (environmental disaster derivate from the Prestige sinking), fishermen have not transferred these effects to prices at port, highlighting its limited market power.

This situation has unleashed an important rentability problem of the inshore subsector. Both the Administration and the sector itself have become aware of this problem. Thus, regarding the demand incentives, the local authorities have been financing advertising campaigns trying to increase the mackerel consumption. As far as the supply is concerned, a pilot experience has been introduced during 2008, consisting on a daily coupon system, per ship and sailor, that limits the fished units in order to induce a rise of the mackerel prices.

Acknowledgments. This work has been financially supported by the assistance programme for training researchers (2007/2010) of the University

of the Basque Country, and the predoctoral fellowship “Ciencias económicas aplicadas al sector medioambiental con especial incidencia en el sector pesquero” of the Basque Government Department of Agriculture, Fishing and Feeding, developed in the Marine Research Unit in the technological centre AZTI-Tecnalia (2006/2007). The author thanks the supervision and direction of Josu Arteche (Department of Applied Economics III, University of the Basque Country) and Arantza Murillas (Marine Research Unit, AZTI-Tecnalia)

References

1. Zivot, E., Andrews, D.W.K.: Further evidence on the great crash, the oil-price shock and the unit-root hypothesis. Journal of Business and Economic Statistics. 10, 251--270 (1992)
2. Dickey, D., Fuller, W.: Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association. 74, 427--431 (1979)
3. Dickey, D., Fuller, W.: Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica. 56, 1057--1071(1981)
4. Kwiatkowski, D., Phillips, P., Schmidt, P., Shin, Y.: Testing the null hypothesis of stationary against the alternative of a unit root: How sure are we that economic time series have a unit root? Journal of Econometrics. 54, 159--78 (1992)
5. Geweke, J., Porter-Hudak, S.: Estimation and applications of long memory time series models. Journal of Time Series Analysis. 14, 221--238 (1983)
6. Engle, R.F., Granger, C.W.J.: Cointegration and error correction representation, estimation and testing. Econometrica. 55, 251--276 (1987)
7. Johansen, S.: Statistical analysis of cointegration vectors. Journal of Economic Dynamics and Control. 12, 231--254 (1988)

Teaching Econometrics with Free Software

Useful software for econometric beginners

Šárka Lejnarová, Adéla Rácková,

University of Economics, Prague, Department of econometrics, Nám. W. Churchilla 4,
Praha 3, 130 67, Czech Republic
Sarka.Lejnarova@vse.cz, Adela.Rackova@vse.cz

Abstract. In this paper we analyzed econometric software suitable for basic econometric lessons. We were limited by software availability at the University of Economics in Prague and that is why the following econometric software was selected for our analysis. We discussed advantages and disadvantages of GiveWin, Gretl and SAS Enterprise Guide and we evaluated them according to their convenience for econometric beginners. Using multiple criteria decision methods, Gretl was chosen as the most suitable econometric software for basic econometric lessons.

Keywords: Gretl, SAS, GiveWin, teaching, basic econometrics

1 Introduction

The importance of the science we all know as econometrics has risen during last years. Econometric lectures were involved in many majors and nowadays not only students with econometric majors but even students with major of economics, finance, national economy, accounting, marketing and management have to deal with this science. Such an expansion of lectures among students who are highly educated in economics but have very basic knowledge of statistics, math and relevant software programs causes many troubles during econometric lessons. The main goal of basic econometric lessons is to introduce students into the ideas of econometric and statistical analysis on real economic dates. This is crucial for the possibility to use a simple econometric analysis in students' final thesis to support their research.

After few lessons we recognized that usage of relevant software during lectures is one of the foundation-stones in the successful education of econometrics. It is not a problem to choose one program and present it to our students but the question is which software is the most appropriate. Majority of students is worried about the software and they concentrate on how to be familiar with the program and the fundamental of the lecture – in the meaning of studying econometrics – is hidden behind this program. That is why we would like to compare used and accessible programs at our university and decide which one is the best or the most acceptable.

At the University of Economics in Prague, we have the possibility to teach econometrics in three programs – in GiveWin, SAS and Gretl. That is the reason why

we chose these three programs for our comparison in this thesis. It is not our goal to make a comparison at the whole market of accessible econometric programs because this market is very wide and the fact is that not all programs are appropriate for students starting with the study of econometrics.

The confrontation of these programs was accomplished on the basis of several different criteria. One of them was the financial availability of the program. We explored if there is any freeware or limited free of charge version of program because we expect that students are not willing to buy the necessary program which they usually need only a half-year of their five years studies. And then we tried to take into account the spread of these programs in the use or in the other lectures.

The other criterion was how the work with the program is intuitive - based on the knowledge of commonly used software like MS Office, and how easy it is to familiarize students with the program in eleven lessons we have during the term. If the program is not complicated, students are more focused on econometrics and then they gain required knowledge of this science.

We compared the heftiness of model creation, the graphic interface of the program and other functions like the graphic or statistic analysis. The important criteria were the ability of the programs to run different data files, their collaboration with MS Office (in the meaning of data export and import) or if the data samples are included in the program or on the program's web pages. We confronted the results provided with the programs and if they are intelligible and well-arranged even for econometric beginners. We focused on whether software displays all statistical tests important for the basic statistical and econometric verification of the model.

In this paper we discussed advantages and disadvantages of mentioned programs and tried to decide which one of these three programs is the most suitable for lessons of econometrics especially from the point of view of econometric students.

2 The illustrative example

A model of dependency of inflation rate on unemployment rate was selected as an illustrative example. This dependency was modeled on data sets from the Czech Republic during time period 1993 – 2007.

The basic Phillips curve describes the inverse relationship between the rate of wage inflation and unemployment rate. We estimated the modified Phillips curve which was published by P. A. Samuelson and R. M. Solow and it represents the inverse relationship between the unemployment rate and inflation rate. Friedman and Phelps criticized the long-term Phillips curve and they proved that the negative slope of the Phillips curve represents the money illusion of the workers and that this illusion is only short-term and there is no tradeoff between the inflation and unemployment and the long-term Phillips curve is vertical (Mach, 2001).

Students have to prove the inverse relationship between inflation and unemployment known as the Phillips curve. Furthermore they have to figure out the goodness of fitted model, test classical linear regression model assumption and then discuss the economic interpretation of the model and model's compliance with the known economic theory.

2.1 GiveWin

GiveWin is the software containing econometric techniques, from single equation methods to advanced cointegration and volatility models. GiveWin contains different program units for different models.

GiveWin was chosen for this analysis because it is the only software used for econometric lessons and modeling of time-series at our university. The interface of the program is similar to the other programs that are familiar to our students.

It is supposed that the user of this program have basic knowledge of statistics to understand the displaying of the output. It is possible to gain an example library with data sets from software websites.

The disadvantage of this software is that there is no full version as freeware and it causes that students can not work with the program on their personal computers without breaking of a copyright. The students can only use a limited student's version.

There is the possibility to create own database in GiveWin but it is not very user-friendly and it is better to export data from other software product although there can be problems with the definition of decimal separators.

The program uses two types of menus. The main menu is the same for all program units and it contains basic operations with data needed for calculations, graphic analysis etc. The second menu is diverse for every program unit and it is used to create the particular model. Switching between main menu of the program and menu in the program unit can be little bit confusing for students.

The standard output involves typical statistical values and tests. The output is divided into two tracks. The first one contains the values for every variable separately and the characteristics of the whole model are included in the second part. The following figure represents the standard output of the program.

```
EQ( 1) Modelling Inflation by OLS (using Data1)
       The estimation sample is: 1993 to 2007

                  Coefficient  Std.Error  t-value  t-prob Part.R^2
Constant              19.2151      3.882     4.95   0.000   0.6534
Unemployment         -2.03631     0.5781    -3.52   0.004   0.4883

sigma                 3.98826  RSS                206.781255
R^2                  0.488309  F(1,13) =     12.41 [0.004]**
log-likelihood       -40.9612  DW                       1.09
no. of observations        15  no. of parameters           2
mean(Inflation)       6.03333  var(Inflation)        26.9409
```

Fig. 1. The output from GiveWin.

The graphical analysis is not involved in this output and it must be done separately. The following figure represents some particular parts of graphical analysis.

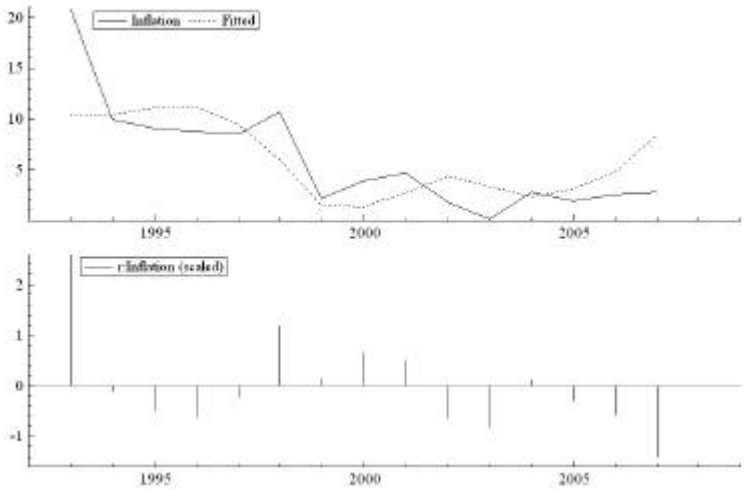


Fig. 2. Graphical analysis in GiveWin.

The graphical analysis is automatically used for predictions. The user has to change this setting if writing results instead of graphing is required.

The evaluation of autocorrelation based on the output can be misleading especially for econometric beginners because only Durbin-Watson statistic is displayed – no matter if it can or can not be used. We discovered problems with entering seasonality into the model.

2.2 Gretl

Gretl is the free and open-source software for econometric analysis with a large database of data sets on web sites.

Gretl was chosen for this analysis because it is downloadable, user-friendly and very intuitive software. Though we did not present this program to our students and we only told them where to download it, they were able to elaborate their half-yearly thesis in this software without further tutorial.

To run a regression, we need to import data files or create a new data file. The program offers a change of data definition during communication with user through simple dialog windows. The file with data does not need to be only in a standard database format and we discovered that Gretl communicates very well with excel and other database files.

The operating with Gretl is similar to the operating with E-views which is the other used statistical and econometric software. This can be an important advantage for future praxis of the students.

The software guides the user through well-arranged and intelligible windows. Each window contains menus which can be used for evaluation, graphic analysis, and advancement of the model or further diagnostic tests on the model.

The basic regression output involves a part with parameters' estimators and variables' characteristics. The characteristics of the model are specified below the first part and appear in the standard format. The figure 3 represents the standard regression output.

```
Model 1: OLS estimates using the 15 observations 1993-2007
Dependent variable: Inflation

               coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const          19,2151       3,88155      4,950    0,0003  ***
  Uneployment    -2,03631      0,578135    -3,522    0,0038  ***

  Mean of dependent variable = 6,03333
  Standard deviation of dep. var. = 5,37264
  Sum of squared residuals = 206,781
  Standard error of the regression = 3,98826
  Unadjusted R-squared = 0,48831
  Adjusted R-squared = 0,44895
  Degrees of freedom = 13
  Durbin-Watson statistic = 1,08853
  First-order autocorrelation coeff. = 0,142501
  Log-likelihood = -40,9612
  Akaike information criterion (AIC) = 85,9223
  Schwarz Bayesian criterion (BIC) = 87,3384
  Hannan-Quinn criterion (HQC) = 85,9072
```

Fig. 3. The output from Gretl.

Students appreciate that the model's characteristics are not named in shortcuts. Using complete names makes the output more intelligible especially to econometric beginners who are not familiar with statistical shortcuts. If necessary for statistical verification, the regression output is extended with additional statistics (e.g. Durbin's *h*) and this is very beneficial and helpful during econometric lessons.

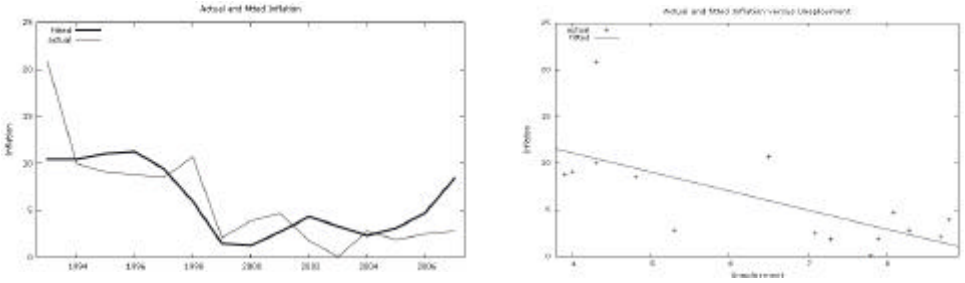


Fig. 4. Graphic analysis in Gretl.

Gretl offers wide range of graphs, two examples are shown on figure 4. This software allows the best editing of graphs among all the compared programs, e.g. saving graphs directly in PDF format.

For econometric analysis and procedures the command-line client program can be used, sample functions and commands are implemented in user's guide.

2.3 SAS Enterprise Guide

SAS Enterprise Guide is a graphical interface of SAS base, which is one of the best known statistical software. SAS Enterprise Guide contains statistical functions, from descriptive statistics to multivariate analysis. SAS Enterprise Guide was chosen for this analysis because it is the software that each student of our university knows from basic statistical lessons. Therefore students are aware of features and settings of SAS Enterprise Guide and they can more concentrate on econometric problems and new functions. The other advantages are the user-friendly interface and the possibility to display the output in html format. This format is more readable for non-statistical users. SAS Enterprise Guide allows writing code in SAS language. Products of SAS Company are widely used in praxes. SAS Enterprise Guide contains an example library with data sets.

A disadvantage of this software is the non econometric character. Also there is no freeware of this software and this causes that students cannot work with it on their personal computers for free.

SAS Enterprise Guide allows you to read datasets in different formats by function *import data*. For running a linear regression there is a simple choice *analyze – regression - linear*. In the window of linear regression you can choose task roles (dependent, explanatory variables), their types (numeric, character), model selection method (stepwise selection), model fit statistics (Akaike's information criterion, Sawa's Bayesian information criterion), additional statistics (confidence interval of parameter estimates on different significance level, partial correlation) and also autocorrelation test and heteroskedasticity test. In the same menu you can choose different types of charts.

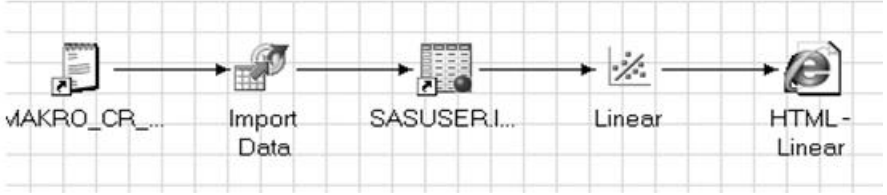


Fig. 5. Project diagram in SAS Enterprise Guide.

The output is automatically displayed in html format and it contains classical statistical test for significance of variables and of model and selected types of plots. This output provides basic information about fulfillment of Gauss-Markov theorem.

In the last tables of output, it is tested whether error terms are independent and identically distributed. Thee last table provides information about autocorrelation and displays Durbin-Watson statistics and autocorrelation coefficient ?.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	197.33208	197.3320	12.41	0.0038
Error	13	206.78126	15.90625		
Corrected Total	14	404.11333			

Root MSE	3.98826	R-Square	0.4883
Dependent Mean	6.03333	Adj R-Sq	0.4489
Coeff Var	66.10382		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	19.21505	3.88155	4.95	0.0003
Unemployement	1	-2.03631	0.57813	-3.52	0.0038

Test of First and Second Moment Specification		
DF	Chi-	Pr > ChiSq
2	3.41	0.1815

Durbin-Watson D	1.089
Number of Observations	15
1st Order Autocorrelation	0.121

Fig. 6. Output from SAS Enterprise Guide.

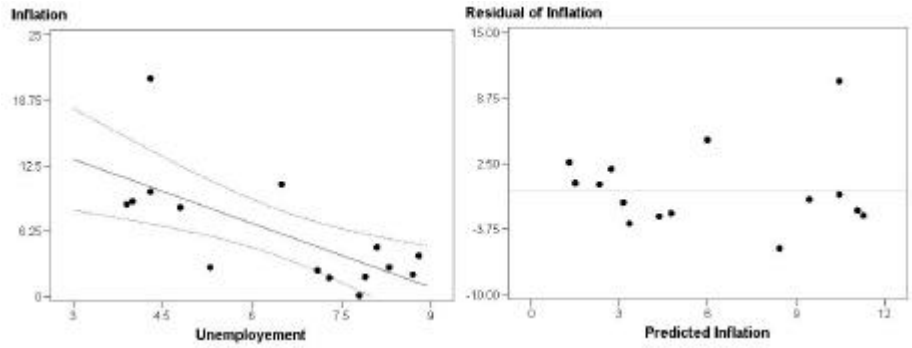


Fig. 7. Residuals in SAS Enterprise Guide.

Working in SAS Enterprise Guide allows you to do a quick and readable analysis. This characteristic is very useful for students working on their final thesis and doing basic analysis of economic variables. For further econometric analysis it is possible to

use other tests by "writing code", samples codes are implemented in SAS Enterprise Guide help.

3 Decision analysis

We applied methods of multiple criteria decision making to choose one of three programs. Each program was evaluated according to seven criteria: Financial availability, Intuitiveness of program, Heftiness of model creation, Graphic interface, Using in praxes, Import of data files, Econometrics function. Methods of multiple criteria decision making are searching a compromise solution by trying "maximize" or "minimize" different criteria. The list of the criteria and the values used in our analysis are in the Table 1.

Table 1. Criteria.

Criteria	Label		GRETL	GiveWin	SAS	Scale
DAT	*Run different data files*	*MAX*	4	3	5	5 - runs different kinds of data files, 0 - just one type of data files
ECO	*Econometrics function*	*MAX*	5	5	3	5 - large scale, 0 - restricted
FIN	*Financial availability*	*MAX*	2	1	0	2 - freeware, 1 - restricted student version, 0 - no freeware
GRA	*Graphic interface*	*MAX*	4	2	5	5 - advanced graphic interface, 0 - restricted graphic interface
HEF	*Heftiness of model creation*	*MAX*	5	2	4	5 - easy fitting, 0 - complecated fitting
INT	*Intuitiveness of program*	*MAX*	5	2	5	5 - user friedly, 0 - user unfriendly
PRA	*Using in praxes*	*MAX*	1	2	5	5 - widly use, 0 - no use

The criteria have a different importance in this analysis. We used Fuller method to define weights of criteria. Fuller method uses the couple comparison of criteria. There is the graphical scheme of comparison called Fuller's triangle.

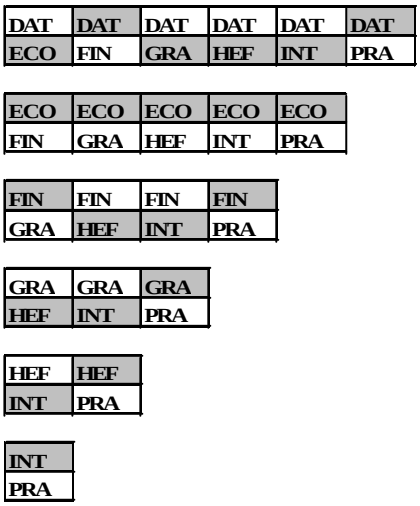


Fig. 8. Fuller's triangle

We computed weights like

$$v_i = \frac{n_i}{\sum_{i=1}^{k} n_i} = \frac{n_i}{N}, \tag{1}$$

where v_i denotes weights of criteria i,
n_i denotes how many times was criteria i selected and
N is total number of comparison obtain by

$$N = \binom{k}{2} = \frac{k(k-1)(k-2)!}{2!(k-2)!} = \frac{k(k-1)}{2}, \tag{2}$$

where k is number of criteria.

The method of weight average was chosen to determinate the compromise solution. Normalized values and weights are in Table 2. The highest weight average has Gretl.

Table 2. Normalized values of criteria and weights.

			Normalized values		
Criteria	**Label**	**Weights**	**GRETL**	**GiveWin**	**SAS**
DAT	*Run different data files*	*0,095*	0,500	0,000	1,000
ECO	*Econometrics function*	*0,286*	1,000	1,000	0,000
FIN	*Financial availability*	*0,095*	1,000	0,500	0,000
GRA	*Graphic interface*	*0,143*	0,667	0,000	1,000
HEF	*Heftiness of model creation*	*0,190*	1,000	0,000	0,667
INT	*Intuitiveness of program*	*0,190*	1,000	0,000	1,000
PRA	*Using in praxes*	*0,000*	0,000	0,250	1,000
Weight average			**0,905**	**0,333**	**0,556**

4 Conclusion

In this paper we compared three different programs used for econometric analysis with the goal to decide which of mentioned software would be the best for basic econometric students at the University of Economics in Prague. We were limited especially with software availability at our university. We decided according to different criteria using multiple criteria decision methods. We discussed the advantages and disadvantages of the programs with respect to usage during our elementary econometric lessons. We did not discuss advanced econometric procedures and methods like reaction functions or panel-data analysis. Those are not a content of basic econometric lectures.

Based on this analysis we recommend using Gretl as the fundamental software for our lectures even if this analysis can not be generalized due to specific conditions as the selection of criteria and the criteria weights used in this paper.

References

1. GRETL webpage: http://gretl.sourceforge.net/
2. Guaranti D.N,: Basic econometrics. McGraw-Hill, Inc.,USA (1988)
3. Hušek, R.: Ekonometrická Analýza. Ekopress, Praha (1999)
4. Jablonský, J.: Operacní výzkum: kvantitativní modely pro ekonomické rozhodování.
5. Professional Publishing, Praha (2007).
6. Mach M.: Makroekonomie II pro magisterské (inženýrské) studium 1. a 2. část, p. 268-269. Melandrium, Slaný (2001)
7. OxMetrix webpage: http://www.oxmetrics.net/
8. SAS webpage: http://www.sas.com/technologies/bi/query_reporting/guide/

Teaching and Learning Econometrics with Gretl

Summarizing Some Experiences

Rigoberto Pérez, Ana Jesús López*

University of Oviedo, Department of Applied Economics,
Campus del Cristo, s/n, 33006, Oviedo, Spain
rigo@uniovi.es, anaj@uniovi.es

Abstract. The European Higher Education Area (EHEA) provides the opportunity of exploring new ways of teaching and learning, emphasizing the role of students in the learning process. In the case of Econometrics (which will play an outstanding role in Economics and Business degrees), our experience has shown that doing Econometrics is a suitable way of learning Econometrics and also that Gretl is a powerful teaching tool, providing our students a wide variety of competences and skills.
This paper summarizes our experiences with Gretl during the last years, including both presential and online learning and combining teachers and students points of view.
Keywords: Econometrics, Competences, Skills, Gretl, AulaNet, Virtual Campus, European Higher Education Area (EHEA)

1 Econometrics in the European Higher Education Area

More than seventy years ago Joseph Schumpeter published his famous work "*The common sense in Econometrics*", where he claimed that every economist is an econometrician since data should be used as a complement of economic theories. Since then, the role of Econometrics in Economics and Business degrees has gradually increased, including not only the study of the main techniques for the estimation and testing of econometric models but also a more realistic approach, which is often based on the use of econometric software.

This more practically-oriented study has become especially important in the present context, since European Universities are currently facing the challenges of the so called "*Bologna process*"[1] which aims to increase the mobility and

* Corresponding author

[1] The Bologna declaration was signed in 1999 by the ministers of education from 29 European countries, with the aim to develop the European higher education area (EHEA) by making academic degree standards and quality assurance standards more comparable and compatible throughout Europe. Since then and after several governmental meetings [Prague (2001), Berlin (2003), Bergen (2005), London (2007)] this process has further developed into a major reform encompassing 45 countries.

employability of European higher education graduates thus ensuring competitiveness of European higher education on the world scale. In the case of Spanish universities, the Bologna process has to face several difficulties since the structure of university degrees in Spain is quite different from the Anglo-Saxon model adopted as a reference[2]

The European dimension of education and the contribution of education in setting the European Information and Knowledge Society have been stressed in the Lisbon Summit (2000) with the strategic goal of "making out of the European Union the world´s most competitive and dynamic knowledge-based economy, capable of sustainable economic growth and with more and better jobs and greater social cohesion". Since Economic and Business degrees are narrowly related to these objectives, in this paper we focus on Econometrics as a strategic tool in these fields, also showing the potential of Gretl in the teaching-learning process and summarizing some recent experiences.

1.1 Competences and skills

A concrete approach to implement the Bologna Process is offered by the project "Tuning Educational Structures in Europe", which provides a methodology to re-design, develop, implement and evaluate studies. Furthermore Tuning serves as a platform for developing reference points at subject area level, which are relevant for making programmes of study comparable and compatible.

According to this approach, reference points are expressed in terms of learning outcomes and competences. Learning outcomes are statements of what a learner is expected to know, understand and be able to demonstrate after a learning experience, while competences represent a dynamic combination of cognitive and meta-cognitive skills, knowledge and understanding, interpersonal, intellectual and practical skills, and ethical values.

Competences can be distinguished in subject specific and generic ones, which can be classified into three types: instrumental competences (cognitive abilities, methodological abilities, technological abilities and linguistic abilities), interpersonal competences (individual abilities like social skills) and systemic competences (combination of understanding, sensibility and knowledge).

In order to identify the most important generic competences, a large scale consultation was organized among graduates, employers and academics. The questionnaire included 30 competences and a total of 101 university depart-

[2] The Spanish system has two kinds of degrees, respectively leading to a medium-level technical profession (three year Diplomatura) and to higher-level professions or academic disciplines (four or five year Licenciatura or Ingeniería). Nevertheless, the Diplomatura has never been the exact equivalent of a BA/BSc, nor the Licenciatura that of a MA/MSC.

ments took place in the consultation, leading to the results[3] summarized in table 1.

Referring to the generic competences, it is necessary to adapt the students´ knowledge and capacities to the labour market requirements, trying to attenuate the traditional existing distance between the perceptions of academics, employers and graduates.

If we focus on the Economics and Business degrees, which is one of the seven areas considered in the Tuning project, the consultation involved 921 graduates, 153 employers and 153 academics, representing a 17% of the total sample.

Although the available information is quite scarce, according to the provisional guidelines these studies should train individuals capable of analyzing and interpreting the functioning of the economy, with the intention of improving the well-being of the society with the achievement of equity and efficiency and in general to approach the analysis of the most relevant economic and social problems.

More specifically, these degrees should provide competences as "*to use analytical instruments in the decision-making processes*" or "*to handle information technologies*", aspects in which Econometrics is expected to play an outstanding role. In fact, the competences related to Econometrics represent a significant proportion of the total list, including some of the items heading the previously described rankings.

In this context, and following the main guidelines of the Bologna process, during the last academic years we have adopted a "*learning by doing*" approach leading to a more realistic methodology, characterized by an intensive use of e-learning and Gretl.

1.2 The role of Gretl in the teaching-learning process

The e-learning, understood as "*the use of new multimedia technologies and the Internet to improve the quality of learning*" has become increasingly popular during the last years. In the University of Oviedo the virtual campus AulaNet was created in 1999, providing a wide variety of online learning facilities for our students and also joining the Shared Virtual Campus of the G9 Group of Universities[4].

[3] The academics were asked to rank seventeen items previously selected from the thirty item list given to graduates and employers. Therefore thirteen items are not present in the academics ranking.

[4] The G9 Group includes nine Spanish public universities: Cantabria, Castilla-La Mancha, Extremadura, Illes Balears, La Rioja, Pública de Navarra, Oviedo, País Vasco and Zaragoza.

Table 1. Rankings of generic competences according to Graduates, Employers and Academics

	Graduates	Employers	Academics
INSTRUMENTAL COMPETENCES			
Capacity for analysis and synthesis	1	3	2
Capacity for organization and planning	10	13	
Basic General Knowledge	20	21	1
Grounding in basic knowledge of the profession	19	23	8
Oral and written communication in your native language	12	11	9
Knowledge of a second language	24	26	15
Elementary computing skills	6	17	16
Information management skills	5	8	
Problem solving	2	4	
Decision- making	15	15	12
INTERPERSONAL COMPETENCES			
Critical and self-critical abilities	17	16	6
Teamwork	14	6	
Interpersonal skills	9	8	14
Ability to work in an interdisciplinary team	21	18	10
Ability to communicate with experts in other fields	18	20	
Appreciation of diversity and multiculturality	29	28	17
Ability to work in an international context	26	27	
Ethical commitment	28	22	13
SYSTEMIC COMPETENCES			
Capacity for applying knowledge in practice	6	2	5
Research skills	25	29	11
Capacity to learn	3	1	3
Capacity to adapt to new situations	8	7	7
Capacity for generating new ideas (creativity)	16	10	4
Leadership	27	25	
Understanding of cultures and customs of other countries	30	30	
Ability to work autonomously	4	12	
Project design and management	23	24	
Initiative and entrepreneurial spirit	22	19	
Concern for quality	11	5	
Will to succeed	13	14	

Since then, the virtual campus AulaNet has experienced some changes, moving from the initial self-developed platform to WebCT and then to Moodle, and providing an increasing number of learning resources, which have been proved to be very useful for communication, teamwork and evaluation.

Furthermore, in the case of Econometrics, our experience shows that the election of suitable software is a main point in order to provide students the required competences and skills. The need of a user-friendly, flexible, open-source and accurate software lead us to Gretl, whose main advantages for teachers and students are described in the next sections.

The outstanding role played by Gretl in the Econometrics teaching-learning process is shown in table 2, which summarizes the main competences related to different learning resources.

Table 2. Teaching-Learning methods in Econometrics and related competences

Teaching-Learning Methods	Competences
Theoretical Sessions	Basic General Knowledge, Capacity for analysis and synthesis, Capacity to learn
Practical Sessions with Gretl	Capacity for applying knowledge in practice, Elementary computing skills, Knowledge of a second language
Team Work with Gretl Database building, Model Specification, Estimation, Testing Analysis, Forecasting , Oral exposition, Final Report	Teamwork, Information management skills, Creativity (capacity for generating new ideas), Problem solving, Decision-making, Capacity to adapt to new situations, Oral and written communication, Concern for quality, Research skills
Self-assessment, Online survey	Ability to work autonomously, Critical and self-critical abilities, Ethical commitment

2 Teaching Econometrics with Gretl

From the teachers´ perspective the use of Gretl offers several advantages, since this open source software provides an easy intuitive interface, allowing different ways of working from interactive point-and-click to batch processing.

This flexibility is one of the most outstanding characteristics of Gretl, offering teachers a good opportunity to re-think contents, methods and evaluation procedures. Therefore, during the last years we have carried out some interesting experiences with Gretl both in presential and online learning.

As it is summarized in table 3 these experiences refer to three different subjects whose characteristics are quite different, but nevertheless we can confirm

that in all the cases the implementation of practical sessions with Gretl has been a successful experience for students and teachers.

Table 3. Description of teaching-learning experiences with Gretl

	Econometrics	**Time Series**	**Forecasting**
Subject Description	Compulsory subject Third year, Degree in Economics, Univ. Oviedo	Optional subject, Third-fourth year, Degree in Economics, Univ. Oviedo	Free-election subject, all years and degrees, G9 Shared Virtual Campus
Teaching-Learning Method	Presential, Blended learning [Theoretical and practical sessions, oral expositions, teamwork]	Presential, Blended learning [Theoretical and practical sessions, oral expositions, personal project]	
Number of students	120	14	Online learning [multimedia facilities, online practical sessions, forum, chat, mail, self-assessment]
Gretl sessions	1 hour/week	2 hours/week	Online sessions
Evaluation	50% Final Exam 30% Teamwork with Gretl 20% Continuous assessment	40% Project with Gretl 20% Oral exposition 20% Continuous assessment 20% Final test	40% Gretl Practices 40% Tests 20% Online participation

It is interesting to stress the role played by Gretl in the whole teaching-learning process, including contents, methods and evaluation. Furthermore, with the aim of achieving coherence, Gretl use has been designed according to the specific syllabus, teaching conditions and student profiles for each of the considered subjects.

Thus, in the case of Econometrics Gretl provides a wide variety of menu options including least squares estimators, maximum likelihood, generalized method of moments, single-equation and system methods, also offering a "*console*" option where users can type commands which are recorded as a batch file.

Our teaching strategy has been designed taking advantage of the flexible character of Gretl. Therefore we combine the quick point-and-click menu options with the use of some commands and console facilities in order to re-build specific results, improving the autonomous work and the comprehension of the main econometric concepts and techniques.

For instance, after obtaining a least squares estimation, students are suggested to obtain the results through matrix computing; some residual tests can be re-build through auxiliar regressions, two stages least squares can be recovered by running the instrumental variables regression and then substituting the estimated results, ...

Nevertheless, the most powerful teaching tool of this subject is the teamwork which provides the students the opportunity to work with real information allowing them to get familiar with the main problems of the econometric modeling. This work is developed with Gretl in 3-4 student groups, also including the presentation of an oral exposition and a final report.

According to our experience, one of the most useful characteristics of Gretl is the session concept, which provides an iconic space containing several objects as data sets, model tables, scalars, graphs, ... thus allowing users to save a complete dossier of the whole developed work.

Regarding Time Series, Gretl facilities include ARIMA and VARMA, VEC, GARCH, unit-root and cointegration tests,... Since this is an optional subject with quite few students, they are required to develop along the semester a personal project consisting in the estimation of time series models for real monthly or quarterly economic data.

Finally, the case of Forecasting shows some differential characteristics referred both to the teaching method and the university context. With regard to the first aspect, while Econometrics and Time Series are based on "*Blended Learning*", understood as "*the combination of different learning styles that can be accomplished through the use of 'blended' virtual and physical resources*", Forecasting is an online free-election matter included in the Shared Virtual Campus of the G9 Group of Universities.

This second aspect must also be stressed due to the diversity of students, belonging to nine Spanish universities and also to different degrees and courses.

Once again, the flexibility of Gretl has been very helpful in the design of this online subject, whose second year will start next February.

In all the described cases, from the teachers perspective we must stress some additional advantages related to Gretl educational and research resources: the access to datasets associated with Econometrics texts as those of Ramanathan, Greene, Stock and Watson, Greene or Gujarati, and the existence of an open scientific community.

3 Learning Econometrics with Gretl

The Bologna process leads to a student-oriented approach, which focuses on the usefulness of study programmes for a future position of the graduate in the

society. Therefore, a sensible definition of learning outcomes and the allocation of credits (according to the European Credit Transfer System, ECTS, based on students´ workload) play a decisive role.

The experiences we are summarizing in this work include students´ participation in these new aspects: the competence-based learning and computation on ECTS credits for different subjects.

Since the virtual campus AulaNet allows an easy implementation of online questionnaires, during the last academic years students have been asked to provide information about their personal effort, the perceived difficulty of the educational contents and the acquired competences and skills. A scheme of the Econometrics survey is shown in table 4.

Table 4. Online survey for Econometrics Students

Sections	Quantitative aspects	Qualitative aspects
Personal Work	Hours of study	Perceived difficulty for each item
Team Work	Hours for different stages: database building, model estimation, hypotheses testing, forecasting, exposition, final report	Perceived difficulty for the team work Satisfaction with Gretl facilities Comparison of personal and team effort Quality of the work compared with others
Quality of the work compared with others Assessmen		Perceived difficulty of assessment questions Satisfaction with the assessment system
General vision		Level of satisfaction with the subject Opinion about professional skills Comments and suggestions

Although the rate of response was quite low (50%), the obtained information shows some interesting facts. A first consideration is the heterogeneity of students (which should be considered when designing and implementing the learning methodology), reflected in the high dispersion of times of personal work, leading to non-representative means.

On the other side a considerable homogeneity is found in the perceived levels of difficulty (approached by the percentages of students whose answers to the proposed questions are "*difficult*" or "*very difficult*") and also in the level

of agreement with the achieved competences and skills (approached by the percentage of students answering "*Total agree*" or "*Agree*"). Besides, since these surveys have been carried out along four academic courses we have also confirmed the stability of the obtained results.

Table 5. Results of the Econometrics online survey

Course Contents	**Level of difficulty**	**Level of agreement**
Econometric Models	0.5%	90,0%
Single Regression Model	17,4%	70,6%
Multiple Regression Model	65,2%	76,9%
Qualitative Variables	43,5%	61,5%
Testing Hypotheses	73,9%	80,0%
Simultaneous Equation Models	50,0%	72,7%
Practical Sessions with Gretl	10,5%	70,0%
Teamwork with Gretl	73,3%	94,1%
Continuous Assessment	17,9%	70,8%
Final Exam	8,3%	100%

Regarding students´ opinions about learning outcomes, they mainly emphasize competences as problem solving, capacity for applying knowledge in practice, computing skills and information management skills. Teamwork is also appreciated as a valuable although rather hard competence.

The application of the described methodology and evaluation system has improved the academic indicators of Econometrics as shown in table 6.

Table 6. Econometrics Academic Indicators

Academic Year	**Proportion of presented students (%)**	**Rate of Success (Proportion of passed students, %)**
2004-05	77%	77%
2005-06	68%	80%
2006-07	73%	80%
2007-08	79%	88%

4 Concluding remarks

In the framework of the Bologna process the new Economics and Business Degrees must face the challenge of training individuals capable of analyzing and interpreting the functioning of the economy improving the well-being of the society with the achievement of equity and efficiency. With this aim, Econometrics should be considered as a strategic tool whose teaching-learning method should be designed paying special attention to competences as problem solving, decision-making or information management.

Since the election of a suitable software is a key point in the teaching-learning process, during the last years we have experienced the potential of Gretl in three different subjects, leading to satisfactory results for both teachers and learners. More specifically, students surveys show the contribution of these learning methods in the achievement of competences as problem solving, capacity for applying knowledge in practice, computing and information management and teamwork.

To end, we would like to thank the colleagues of the G9 Virtual Campus for sharing valuable experiences during nine years, our students for their collaboration in the online surveys and the Gretl Community for developing and spreading the use of open source econometric software.

Bibliography

[1] Commission of the European Communities: Challenges for the European Information Society beyond 2005, Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions, 757 (2004)

[2] Commission of the European Communities.: Progress towards the Lisbon objectives in Education and Training, Commission Staff Working Paper, 419 (2005)

[3] Cottrell, A.; Lucchetti, R.J.: Gretl User´s Guide, Gnu Rregression, Econometrics and Time Series (2008)

[4] Esteban, M.V. et al: Análisis de regresión con Gretl, Departamento de economía Aplicada III, Universidad del País Vasco.

[5] González, J.; Wagenaar, R.: Tuning Educational Structures in Europe. Universities´ contribution to the Bologna Process, http://tuning.unideusto.org/tuningeu (2005)

[6] González, J.; Wagenaar, R.: Universities´contribution to the Bologna Process, http://tinyurl.com/anylvc (2008)

[7] López, A.J.: El papel del E-Learning en el Espacio Europeo de Educación Superior, Congreso Online Educa, Madrid (2004)

[8] López, A.J.; Pérez, R.: An experience on virtual teaching: AulaNet, in Computers and Education: Towards an Interconnected Society, M. Ortega and J. Bravo Ed., Kluwer Academic Publishers, pp. 207—214 (2001)

[9] López, A.J.; Pérez, R.: Networking Universities to bridge the Digital Divide, International Journal of Instructional Technology and Distance Learning, vol.3, n.5, pp. 73–82. (2006)

[10] López, A.J.; Pérez, R.; Mayor, M.: Learning Econometrics by Doing Econometrics. Some pilot experiences, ISI Lisboa (2007)

[11] Pagani, R.; González, J.: El crédito europeo y el sistema educativo español, Informe Técnico ECTS Counsellors and Diploma Supplement Promoters (2002)

[12] Reichert, S.; Tauch, C.: Bologna four years after: Steps toward sustainable reform of higher education in Europe", Trends 2003: Progress towards the European Higher Education Area, European University Association (2003)

[13] Schumpeter, J.: The common sense in Econometrics, Econometrica, Vol. 1, n.1, p. 5–12 (1933)

The Roster-in-a-Box Course Management System

Tavis Barr

Long Island Unversity/Makerere University

Abstract. Roster-in-a-Box is an open source course management system written in PHP, with a MySQL back end. It is designed to handle only the homework part of course management, leaving the instructor to design other parts of the course web page. It includes a complete semester of auto-graded assignment modules for introductory and intermediate microeconomics, and for introductory statistics, which I have quality-tested by using multiple times in my courses. I have evaluated the effectiveness of the system by comparing student test scores in my introductory statistics course before and after I began using the system (the material in the other courses changed to much before and after for exam performance to be comparable); there is weak evidence of improvement, and no evidence of worse student performance. I briefly discuss ways in which the software might be improved.

1 Introduction

Roster-in-a-Box is a course management system designed to facilitate the use of autograded homework assignments, while also allowing for text-based questions to be submitted online and graded online by the instructor. I developed and currently use the system for my introductory statistics and undergraduate microeconomics courses at Long Island University, and I have had it in production since 2005. While any instructor who tries hard enough could break the program, it is stable and very much ready for use.

I wrote the program because I wanted something simpler than Moodle (and also because my school was not supporting Moodle at the time that I started the project). I have no complaints about Moodle, but it was not right for me, because I did not want the course management system to take over my course web site. Roster-in-a-Box handles the homework and grading functions with a couple of web pages that can be inserted into any course web site. It requires the web server to run PHP and MySQL, but otherwise can be used quite flexibly.

The program consists of two web pages, namely a homework page and an administration module, as well as several utilities. Each page is its own PHP program. The homework page can be linked to from the main course web page, and is responsible for displaying, correcting, and recording all homework assignments. The administrative module, though password protected, should not

be made viewable to the rest of the world; the instructor can use this module to set up and modify the list of homework assigments, as well as to grade the manually graded assignments and track student performance. Additionally, small utilities display web pages through which students can request an account and view their grades.

The package comes with a set of modules for three different courses: Introductory microeconomics, intermediate microeconomics, and introductory statistics. I have used all of these modules in my teaching, so they are tested in that way. Each set consists of around thirty to fifty autograded questions based on material from the respective course; the instructor merely needs to use the administrative page to create an assignment, insert a module or modules into that assignment, and set the number of points for each question. The software takes care of the rest, although manually graded text questions can also be given. When a student enters an incorrect answer, the homework module explains how to solve the problem correctly. The student has as many chance as she wishes to try the assignment again with a different set of problems. Instructors can set a minimum score requirement below which the system will not accept an assignment. It is also possible to provide a late penalty for every day beyond the due date that the assignment is submitted, as well as a final cutoff date beyond which the system will not accept assignments.

The statistic modules cover summary statistics, probability, random variables, the Central Limit Theorem, confidence intervals, one- and two-sample hypothesis tests, and univariate regression. Nearly all of the sample problems are based on real-world data, and contain a sufficient number of exercises that it is very unlikely for a student to see the same problem twice. The introductory microeconomics modules cover topics such as supply and demand, consumer choice and demand, market demand, production, monopoly power, perfect competition, oligopoly, public goods, and externalities at the US introductory level. The intermediate microeconomics modules cover the same material at the US intermediate level, and are meant to accompany my open-source textbook, Intermediate Microeconomics, available from my web site at http://myweb.liu.edu/~tbarr. To some degree, the exercises reflect my idiosyncratic way of presenting the material for these courses, but I am confident that every instructor will find useful questions from the modules, and I welcome contributions from others.

There are also two administrative modules that I found a need for: One e-mails students when they have overdue assignments, leaving them without the excuse to complain that they did not know an assignment was due. It is also possible to e-mail announcements to the class using the main administrative page. The other module e-mails a backup of the roster to the administrator and can be

set to run on a daily or weekly basis; aside from protecting against catastrophic system failure, I find this useful when I need to respond to students who claim that the system "lost" their assignment.

2 Teaching Experience

Although I have used Roster-in-a-Box in three courses, I can only conduct a plausible before-and-after comparison in my introductory statistics course, because the material in the other two courses (introductory and intermediate microeconomics) changed as I implemented the system – chiefly because the time saved in class by not going over homework assignments freed me up to cover more material. Therefore I have examined the quantitative determinants of student performance in the statistics class, and their changes over time, to gauge effectiveness of the course management software.

Since the Summer of 2004, I have used the same computer program to generate my midterms and final exams. The program allows each student to get an exam with identically-worded questions but different answers, thereby making it far more difficult for students to cheat by copying answers from their neighbors. I change the wordings of the questions every year, but not the statistical calculations involved; this makes exam scores comparable year after year. In Summer 2004, the midterm covered measures of central tendency and dispersion, probability theory, and discrete random variables, while the final exam covered the Normal distribution, the Central Limit Theorem, confidence intervals, and one- and two-sample hypothesis tests.

The only major change came in 2005, when I switched to an online homework assignment system. Because I no longer had to go over homework assignments in class, my lecture time was freed up to cover more material. I therefore added one more topic (the Normal distribution) to the midterm, and added univariate linear regression to the final exam in its place. Therefore, the midterm became a bit harder (since it covered one more topic), while it is difficult to say whether the final exam became harder or easier (my inclination is to think that it became harder because regression is more difficult than the Normal distribution, but this is certainly a matter of perspective).

Other factors may have made exam performance incomparable between semesters. It is possible that my ability to explain material has improved, for example. I have also replaced contrived examples during my lectures with examples from actual data sources, which may keep students more interested. I do more in-class problem solving than I used to. And, I devote a little bit more attention to following up with under-performing students than I used to. Nevertheless, aside from the increase in material, the course content and the level

of difficulty has pretty much remained constant. It is with all of these caveats that I am comparing midterm and final exam scores between semesters over the last three and a half years. The homework scores are certainly not comparable before and after I introduced the online homework system, if nothing else because I used a different marking system. Nevertheless, within each of the two homework regimes, the assignments are quite similar.

Figures 1 and 2 show the mean, 25th, and 75th percentiles of exam performance between Summer 2004 and mid-Fall 2007 (Spring and Fall 2006 are missing because I was on leave). The online grading system was introduced in semester 4 (Summer 2005). The means suggest a slight upward trend, though with as much random variation between semesters as over time. The 75th percentile shows relatively little movement, while the 25th percentile is quite erratic; I can only hope that the apparent upward movement of that percentile over the last two to three semesters will continue.

Table 1 shows t-tests of whether the exam performance was better before or after the online grading system was introduced. The point estimates suggest a mean improvement of about four percentage points in both the midterm and the final exam, however neither result is statistically significant as a two-tailed test (both are borderline significant when considered as a one-tailed test). It is worth bearing in mind that the exams in the "after" case cover slightly more material.

Table 2 shows a regression of midterm performance on homework performance before and after online grading was implemented. The explanatory power of the homework exercises drops slightly after the implementation of online grading. I ascribe this mainly to the fact that the first homework assignment became easier (which allows the students a chance to adjust to the online grading system), and therefore a poorer predictor of midterm performance; the less significant coefficient on this assignment bears out that hypothesis. However, it is also possible that before I implemented the online system, I gave harsh marks to students whose homework pages suggested a poor understanding, holding constant the number of questions the student got correct, which would have made manual grading a more effective way to predict performance based on subjective indications. Table 3 repeats the same exercise for final exam performance. Here, the level of explanatory power is not too different between the online and manual grading regimes.

Table 4 performs regressions of final exam performance on both the relevant homework assignments, as well as on the first four homework assignments and midterm performance. The latter items should not contain any relevant material, but may predict overall student performance. In fact, these indicators taken together can predict about two thirds in the variation of final exam performance both before and after the online grading system was implemented. Table 5 shows

that the midterm score is a less successful predictor of the final exam score in more recent semesters; I hope that this is because I have been increasingly vigilant about intervening in the cases of students who perform poorly on the midterm.

Finally, Table 6 compares performance across different types of semesters. As my intuition would tell me, students in the Summer semester perform significantly higher than academic-year students (the point estimate is ten points higher, and the result is significant at the one percent level), but there is no significant difference between Spring and Fall semesters.

Overall, these numbers suggest that the online grading system had been modestly successful at improving student performance. Exam scores appear to have improved slightly and certainly did not worsen, and moreover, those exams cover more material than before the online system was implemented. The bottom quartile of student performance has not improved as much as I would like, and I should work on developing better mechanisms to assist the performance of this particular group of students.

3 Future Directions

Although I have a professional background in programming and systems administration, I wrote Roster-in-a-Box as a teacher trying to get a particular job done for myself, rather than as a developer generating a product for a particular audience. Nevertheless, I put a certain amount of effort into ensuring that the code is modular, generic, and reusable, and I hope that others may find that Roster-in-a-Box fits their needs.

At this point, the software does what I need it to do, so any further development would be to improve its usefulness to others. I have two ideas in mind. First, although the software does not take over an entire course web site, some instructors would prefer a product that does – that is, they should merely need to enter in topics on a syllabus template, links to course readings, etc., and the course web site is ready to go without them ever having to touch any HTML. I may try to add a module that allows users to do this.

The project that I would find most personally useful is to improve and expand the existing homework modules. I would like to add more questions for all of the courses; I would also like to add modules for other standard economics courses, such as macroeconomics, finance, industrial organization, health economics, etc. Finally, I would like to add features to the modules: First, by using AJAX routines so that students can draw graphs online; and second, allowing students to submit arbitrary files (such as spreadsheets, word processing documents, etc.) to be graded and returned.

Because modules can be written one-by-one, I hope that I can interest others who use the software to write a module or two on their own, and then send it back to me to include in the software. First, however, this requires getting others to actually use the software. I would urge readers and participants to ask themselves: What is the main reason you would not wish to use Roster-in-a-Box? What improvements in the software, if any, would get you to want to use it? I welcome any frank and honest feedback and suggestions.

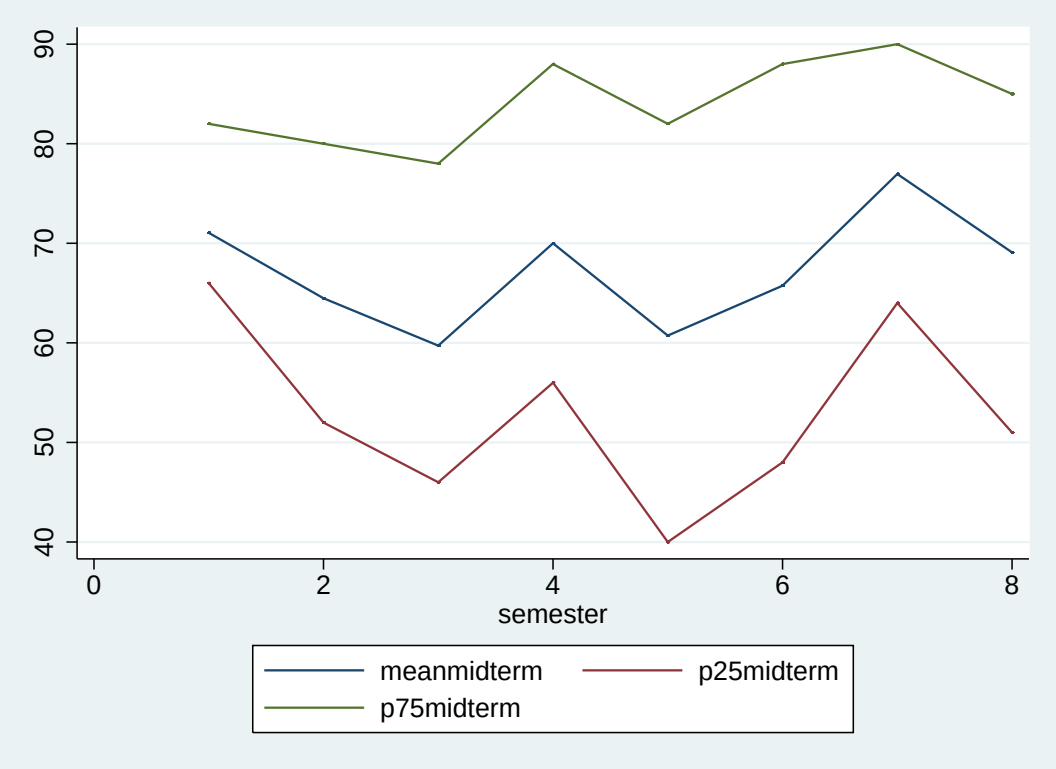


Fig. 1. Midterm Exam Performance by Semester

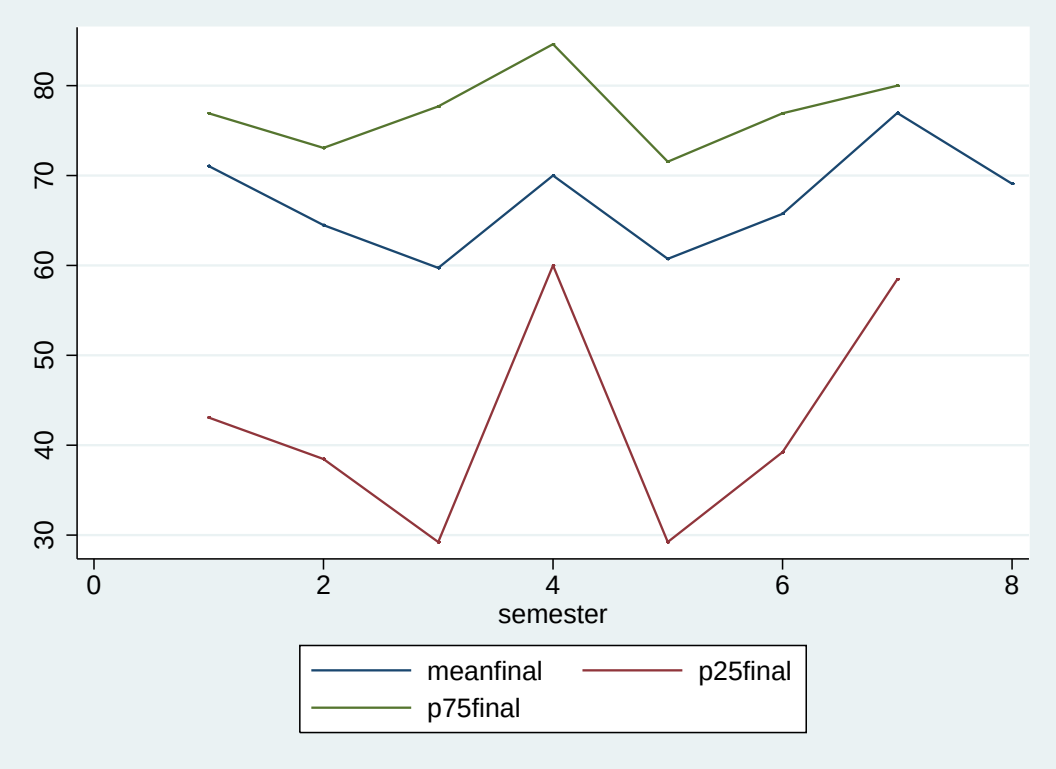


Fig. 2. Final Exam Performance by Semester

Exam	**No. Obs**	**Mean**	**Std. Dev.**	**95% Confidence Interval**
Midterm, Before Online Grading	89	$\bar{x}_1$ =63.67	21.03	(59.24,68.10)
Midterm, After Online Grading	204	$\bar{x}_2$ =67.02	22.64	(63.89.70.14)
Final, Before Online Grading	87	$\bar{x}_3$ =54.05	23.47	(49.05,59.05)
Final, After Online Grading	140	$\bar{x}_4$ =58.22	24.65	(54.10,62.34)

t test, equal variances assumed	**t Statistic**	**p-value**
$\mu_1 = \mu_2$ vs.$\mu_1 \neq \mu_2$	-1.1881	0.2358
$\mu_1 = \mu_2$ vs.$\mu_1 < \mu_2$	-1.1881	0.1179
$\mu_3 = \mu_4$ vs.$\mu_3 \neq \mu_4$	-1.2616	0.2084
$\mu_3 = \mu_4$ vs.$\mu_3 < \mu_4$	-1.2616	0.1042

Table 1. Means of midterms and finals, before vs after online grading

Independent Variable	**(1) Before Online Grading**		**(2) After Online Grading**	
Homework #1 Score (Old)	6.9791***	(1.8372)		
Homework #2 Score (Old)	7.9246***	(1.8897)		
Homework #3 Score (Old)	5.7744***	(1.6806)		
Homework #4 Score (Old)	5.8010***	(1.8763)		
Homework #1 Score (New)			0.1611**	(0.0768)
Homework #2 Score (New)			0.2107***	(0.0764)
Homework #3 Score (New)			0.2291***	(0.0678)
Homework #4 Score (New)			0.1898***	(0.0733)
No. Assignments Submitted	-20.7491***	(5.5652)	-11.2337**	(4.7705)
Summer '04	10.1041**	(5.0461)		
Fall '04	3.0874	(4.1672)		
Summer '05			1.5457	(6.6153)
Fall '05			-7.7052**	(3.7639)
Spring '07			-2.4503	(3.6773)
Summer '07			2.6072	(4.7630)
R^2	0.4095		0.2923	
Obs	89		204	

Standard errors in parentheses. Asymptotic two-tailed significance levels: * 10 percent; ** 5 percent; *** 1 percent

Table 2. Regression of midterm exam performance on homework, (1) before online grading, (2) after online grading

Dependent Variable	**(1) Before Online Grading**		**(2) After Online Grading**	
Homework #5 Score (Old)	7.2570***	(1.7980)		
Homework #6 Score (Old)	8.7313***	(1.6275)		
Homework #7 Score (Old)	7.9676***	(1.6827)		
Homework #8 Score (Old)	3.9323**	(1.7949)		
Homework #5 Score (New)			0.2057*	(0.1132)
Homework #6 Score (New)			0.3284***	(0.1190)
Homework #7 Score (New)			0.1288	(0.0861)
Homework #8 Score (New)			0.1192*	(0.0682)
No. Homeworks Submitted	-20.0672***	(4.9405)	-7.4453	(4.9486)
Summer '04	11.5864**	(5.8476)		
Fall '04	5.6321	(5.4789)		
Summer '05			12.6164**	(5.9735)
Fall '05			-2.9902	(3.5281)
Summer '07			7.3281*	(4.3555)
R^2	0.5213		0.5081	
Obs	87		140	

Standard errors in parentheses. Asymptotic two-tailed significance levels: * 10 percent; ** 5 percent; *** 1 percent

Table 3. Regression of final exam performance on relevant homework, (1) before online grading (2) after online grading

Independent Variable	**(1) Before Online Grading**		**(2) After Online Grading**	
Homework #1 (Old)	3.1362*	(1.8240)		
Homework #2 (Old)	1.8947	(1.7970)		
Homework #3 (Old)	0.4457	(1.6675)		
Homework #4 (Old)	4.0903**	(1.9613)		
Homework #5 (Old)	3.1358*	(1.6517)		
Homework #6 (Old)	3.5962**	(1.6867)		
Homework #7 (Old)	4.6674**	(1.8224)		
Homework #8 (Old)	1.1219	(1.6662)		
Homework #1 (New)			-0.0305	(0.0812)
Homework #2 (New)			-0.0561	(0.0901)
Homework #3 (New)			-0.0026	(0.0723)
Homework #4 (New)			-0.1029	(0.0766)
Homework #5 (New)			0.1116	(0.1035)
Homework #6 (New)			0.2813***	(0.1043)
Homework #7 (New)			0.1099	(0.0768)
Homework #8 (New)			0.0839	(0.0707)
No. Submitted, 1st Half	-12.4215**	(5.4420)	3.2414	(5.6572)
No. Submitted, 2nd Half	-6.7012	(4.9813)	-4.8447	(4.8897)
Midterm Score	0.5373***	(0.0948)	0.4798***	(0.0700)
Summer '04	6.5092	(4.9461)		
Fall '04	1.3720	(4.4679)		
Summer '05			9.8884*	(5.2891)
Fall '05			-1.3470	(3.1116)
Summer '07			3.9309	(3.7844)
R^2	0.7215		0.6528	
Obs	86		138	

Standard errors in parentheses. Asymptotic two-tailed significance levels: * 10 percent; ** 5 percent; *** 1 percent

Table 4. Regression of final exam performance on all homework, midterm

	Correlation	No. Obs
Before Online Grading	0.7288	86
After Online Grading	0.6612	138

Table 5. Correlation of midterm exam performance with final exam performance

Semester	**No. Obs**	**Mean**	**Std. Dev.**	**95% Confidence Interval**
Summer	89	$\bar{x}_1$ =73.57	18.70	(68.47,78.68)
Spring or Fall	204	$\bar{x}_2$ =64.29	22.58	(61.46,67.17)
Spring	87	$\bar{x}_3$ =63.29	23.77	(58.52,68.05)
Fall	140	$\bar{x}_4$ =64.99	21.78	(61.36,68.62)

t test, equal variances assumed	**t Statistic**	**p-value**
$\mu_1 = \mu_2$ vs.$\mu_1 \neq \mu_2$	-2.8092	0.0053
$\mu_1 = \mu_2$ vs.$\mu_1 > \mu_2$	-2.8092	0.0027
$\mu_3 = \mu_4$ vs.$\mu_3 \neq \mu_4$	-0.5742	0.5663
$\mu_3 = \mu_4$ vs.$\mu_3 > \mu_4$	-0.5742	0.2832

Table 6. Comparison of Performance Across Semesters

Contributions to gretl Development

On Embedding Gretl in a Python Module

Christine Choirat and Raffaello Seri

[1] Department of Quantitative Methods,
School of Economics and Business Management,
Universidad de Navarra, Pamplona (Spain)
cchoirat@unav.es
[2] Dipartimento di Economia,
Università degli Studi dell'Insubria , Varese (Italy)
raffaello.seri@uninsubria.it

Abstract. Additional functionalities can be developed for Gretl either directly in the main C code or with the Gretl scripting language. We illustrate through an example how it would be possible to wrap the C source of Gretl with SWIG to create an interface to Python that makes use of the matrix library NumPy. Such an interface would make it easier for users to extend Gretl since it would allow for developing and distributing Gretl extensions as Python modules.

Key words: Python, C, SWIG, Libgretl API

1 Introduction and motivation

To extend the functionalities of Gretl,[3] it is either possible to add them in the C source or to use the Gretl scripting language. The former option is only possible for users who have a very sound knowledge of C and who understand how the source of Gretl is structured. Besides, it makes it hard to share these added functionalities with the rest of the user base unless they are accepted for a next release. The latter option requires learning yet another field-specific language. Even if the speed benchmarks are good and the syntax easy, a program-specific macro language can never be as powerful as a full-featured scripting language (either domain-specific such as R, see [1, 2], or general such as Python, see [3]).

The scripting language that we have chosen to embed Gretl in is Python, which is free, open-source and available on many platforms. Python is indeed a very powerful, easy-to-learn and well-documented language with a very clear syntax. Writing and distributing documented Python modules is simple as well. Matrices are not a native Python data type. However, the very mature project NumPy[4] provides an efficient implementation of $N-$dimensional arrays (and

[3] See http://gretl.sourceforge.net/.
[4] See http://www.scipy.org/.

therefore matrices) accessible via its C API. Moreover, many other Python modules are of interest, such as the NumPy-based scientific library SciPy[5] or the plotting module matplotlib.[6] So, the objective is to get the Gretl embedding make use of NumPy.

In Section 2, we review the tools that can be used to embed C code in a Python module. We show that SWIG is the most powerful one (though not the easiest). Then in Section 3, we see how Gretl could be embedded. Section 4 concludes.

2 Possible implementation choices

Let us consider the following code (which is a simplified version of the f2py example found on the SciPy wiki).[7]

```
/* example1.c */
void func(int n, double *x) {
  int i;
  for (i=0; i<n; i++) {
    x[i] = x[i] + i;
  }
}
```

Python cannot be extended as trivially as R (in particular as can be done through the `.C` function, see Section 5.2 of the R manual [4]) and some nontrivial knowledge of the NumPy C API[8] is required even in the case of such a simple function (see *e.g.* Chapter 14 of [5]). Fortunately, two popular tools, namely f2py and SWIG, make extension development much easier.

2.1 f2py

f2py[9] is a utility that is primarily meant to wrap Fortran code as Python/NumPy modules. However, it seems to be the easiest way to wrap C code as well. It generates a documented Python module that works very smoothly with NumPy. The only required step is to write a signature file, which states that the function to be wrapped is a C function and provides the necessary information about its arguments:

[5] SciPy provides among others routines for statistics, optimization, numerical integration, linear algebra, Fourier transforms, signal processing, image processing, genetic algorithms, ODE solvers and special functions.

[6] See `http://matplotlib.sourceforge.net/`.

[7] See `http://www.scipy.org/Cookbook/f2py_and_NumPy`.

[8] See `http://docs.scipy.org/doc/numpy/reference/c-api.html`.

[9] See `http://www.scipy.org/F2py`.

```
! m1.pyf
python module m1
interface
  subroutine func(n,x)
    intent(c) func
    intent(c)
    integer intent(hide), depend(x) :: n=len(x)
    double precision intent(inplace) :: x(n)
  end subroutine func
end interface
end python module m1
```

Compilation is straightforward: `f2py m1.pyf example1.c -c` generates a Python extension module `m1.so` (on Linux, since the extension is platform-dependent). On the Python side, we get (the »> symbol indicates the Python prompt):

```
»> import numpy, m1
»> a = numpy.array([1., 3., 5.])
»> a
array([1., 3., 5.])
»> m1.func(a)
»> a
array([ 1.,  4.,  7.])
»> print m1.func.__doc__
func - Function signature:
  func(x)
Required arguments:
  x :  rank-1 array('d') with bounds (n)
```

However, as far as we know, no large-scale project as ever been wrapped with f2py, since a lot of manual work is required: an interface has to be written for every C function.

2.2 SWIG

On the other hand, SWIG[10] allows for an almost automatic wrapper generation (including such things as documentation generation and exception handling), the price to pay however is that integration with NumPy is not as easy as with f2py by default. SWIG stands for "Simplified Wrapper and Interface Generator"

[10] See `http://www.swig.org/`.

and allows for interfacing C (and C++) code with several target languages (C#, Java, Ocaml, Octave, Python, R, to name a few). It is successfully used in very large scale projects, one of the most famous being wxPython,[11] which provides Python bindings to the C++ cross-platform GUI library wxWidgets.[12]

For standard data types (`int`, `double`, ...), SWIG works very smoothly. Let us consider the following case, where we have a C file and its associated header file.

```
/* example2.c */
#include "example2.h"
int add1(int n) {
    return n + 1;
}
double add2(double x, double y) {
    return x + y;
}
```

and

```
/* example2.h */
int add1(int n);
double add2(double x, double y);
```

The only required step is to write an interface file `example2.i` (made of two main parts, the headers to be directly included and the functions to be parsed).

```
// example2.i
%module m2
%{
#define SWIG_FILE_WITH_INIT
// Includes the header in the wrapper code
#include "example2.h"
%}
// Parses the header file to generate wrappers
%include "example2.h"
```

and use SWIG to generate wrappers for the target language Python:

```
swig -python example2.i
```

[11] See `http://www.wxpython.org/`.
[12] See `http://wxwidgets.org/`.

A simple Python script

```
# setup_example2.py
from distutils.core import setup, Extension
setup(name="m2",
      ext_modules=[Extension("_m2",
                             ["example2.i", "example2.c"])])
```

takes care of building the extension module `m2`:

```
python setup_example2.py build_ext --inplace
```

From the Python interpreter, we get:

```
»> import m2
»> m2.add1(3)
4
»> m2.add2(3, 4.0)
7.0
```

2.3 SWIG and NumPy

Applying directly the method of Section 2.2 to `example1.c` (and the associated header `example1.h`) leads to a Python function that does not behave the expected way. If we build a module `m3` as above, we get:

```
»> import numpy, m3
»> a = numpy.array([1., 3., 5.])
»> m3.func(a)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: func() takes exactly 2 arguments (1 given)
»> m3.func(3, a)
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: in method 'func', argument 2 of type 'double *'
```

The problem comes from the fact that SWIG does not know how to transform pointers into NumPy arrays. However, it is possible to define *typemaps*, which are SWIG procedures written in a C-like syntax, in order to make the transformation easier. Many of these typemaps are already available in the file `numpy.i` (which is part of the NumPy distribution). We can use it in the interface file `example1.i` to specify (with `%apply`) that the arguments of the function call are the length of the array and the array itself:

```
// example1.i
%module m4
%{
#define SWIG_FILE_WITH_INIT
#include "example1.h"
%}
%include "numpy.i"
%init %{
import_array();
%}
%apply (int DIM1, double* INPLACE_ARRAY1) {(int n, double *x)};
%feature("autodoc", "func(array x) -> x") func;
void func(int n, double *x);
```

We now get the expected behavior, together with a customized documentation (that could also be generated automatically):

```
»> import numpy, m4
»> a = numpy.array([1., 3., 5.])
»> m4.func(a)
»> a
array([ 1.,  4.,  7.])
»> print m4.func.__doc__
func(array x) -> x
```

3 SWIG and Gretl

As we have seen in Section 2, SWIG can be used to wrap a program as large as Gretl in an almost automatic way. Let us consider the minimalistic (and obviously very incomplete) case of three interface files (assuming the source code of Gretl is in `src/`), namely `libgretl.i`, `version.i` and `gretl.i`

```
// libgretl.i
%module gretl
%{
#include "src/libgretl.h"
%}
%include "src/libgretl.h"

// version.i
%module gretl
```

```
%{
#include "src/version.h"
%}
%include "src/version.h"

// gretl.i
%module gretl
%include "version.i"
%include "libgretl.i"
%{
#define SWIG_FILE_WITH_INIT
%}
```

The following setup script (hard-coding the directory of the XML library) generates a Python module `gretl`:

```
# setup_gretl.py
from distutils.core import setup, Extension
setup(name="gretl",
      ext_modules=[Extension("_gretl", ["gretl.i"],
      include_dirs=["/usr/include/libxml2/"])])
```

The module can then be called from Python:

```
»> import gretl
»> gretl.GRETL_VERSION
'1.8.0'
```

C structures are automatically transformed into Python classes. For example, the following structure (in `src/libgretl.h`)

```
typedef struct _cmplx cmplx;
struct _cmplx {
    double r;
    double i;
};
```

can be accessed from Python as

```
»> gretl._cmplx
<class 'gretl._cmplx'>
»> z = gretl._cmplx()
```

```
»> z.r
0.0
»> z.i
0.0
```

4 Conclusion and further developments

We have provided an illustration of the fact that it is possible to embed Gretl's functionalities into a high-level scripting language such as Python. At this point, it would be very fruitful to get some comments from the Gretl community.

- Which target language should be focused upon? Python is the easiest in terms of SWIG integration, but R has more built-in econometric functions and is of more widespread use. (Remark however that the Python package RPy[13] allows for using R from within Python. Moreover, the R community tends now to favor the use of the Rcpp[14] package to create bindings to large C/C++ libraries such as RQuantLib[15]).
- How far should we go in the use of SWIG? The source code of Gretl is large and complex. Gretl defines its own matrix library and matrix operations (in `src/gretl_matrix.c`), its own optimization functions (in `src/gretl_bfgs.c`), etc. All these tools (and many more) are available through NumPy (not to mention Python's XML standard library and the graphical library matplotlib). So, some help from the Gretl developers would allow for selecting the crucial parts to be wrapped.

Writing the proper SWIG typemaps (along the lines of the C++ machine learning library Shogun,[16] see [6]), it would be possible to build a Gretl module for R, Python, Octave and Matlab (and any other language supported by SWIG) with only a minimal effort.

[13] See `http://rpy.sourceforge.net/`.
[14] See `http://cran.r-project.org/web/packages/Rcpp/index.html`.
[15] See `http://cran.r-project.org/web/packages/RQuantLib/index.html`.
[16] See `http://www.shogun-toolbox.org/`.

Bibliography

[1] Cribari-Neto, F., Zarkos, S.: R: yet another econometric programming environment. Journal of Applied Econometrics **14**(3) (June 1999) 319–329

[2] Racine, J., Hyndman, R.: Using R to teach econometrics. Journal of Applied Econometrics **17**(2) (April 2002) 175 – 189

[3] Choirat, C., Seri, R.: Econometrics with Python. Journal of Applied Econometrics (forthcoming)

[4] R Development Core Team: Writing R extensions. `http://cran.r-project.org/doc/manuals/R-exts.html`

[5] Oliphant, T.: Guide to NumPy. `http://numpy.scipy.org/numpybook.pdf`

[6] Sonnenburg, S., Raetsch, G., Schaefer, C., Schoelkopf, B.: Large scale multiple kernel learning. Journal of Machine Learning Research **7** (July 2006) 1531–1565

An Alternative to Represent Time Series: "the Time Scatter Plot"

Alberto Calderero[1], Hanna Kuittinen[1], and Javier Fernández-Macho[2]

[1] Innovation Systems Unit, LABEIN-TECNALIA
acalderero@labein.es
[2] Department of Applied Economics III, University of the Basque Country

Abstract. This paper presents a substantial improvement for the representation of panel data. The time scatter plot is an alternative type of graph using Cartesian coordinates to display values across the time for two variables of a set of data. The data is displayed as a collection of points, linked by lines and ending by an arrow, each having the value of one variable determining the position on the horizontal axis and the value of another variable determining the position on the vertical axis. The purpose of the lines and the final arrow is to link the consecutive points in time for each case, resulting in a chain of lines in the shape of an irregular arrow which starts in the first period of time and ends in the last. This paper illustrates the usability of time scatter plot with two case studies 1) Research and development as a driver of innovation and growth and 2) The goal of social and economy convergence in Europe.

Key words: time scatter plot, convergence plot, graphing time series, plotting tools.

1 Introduction

Graphical demonstration is a very powerful means to present statistical data. For presenting a long-term development in time or relationship between two variables a graph can compress the data in a very interpretable manner for the use of decision support (e.g. policy makers, managers of companies). As the end-users of statistics are not necessarily very familiar with statistical models, a well design and structured graph can give the requested information in a form that is very easy to understand.

This paper presents a substantial improvement from the deficient present graphical representation of panel data (in Gretl and probably in other statistical packages). Gretl's graphical representation of panel data simply shows each of the "groups"=cases sequentially (with the option of individual graphs if the number of cases is small), which is not satisfactory when the number of cases is relatively high. Besides, this is only useful to present the time evolution of some cases, whilst there is nothing to visualize the evolution of the relationship between two characteristics of the whole set of cases in the panel, which is precisely what the present instrument does.

Our graphical proposal shows an alternative way to represent time series in a scatter plot. When compared to classical scatter plot, the advantage of so called time scatter plot is in its representation capacity. It can show two variables and their evolvement in time for several cases (or individuals) using a simple variation of a traditional scatter plot.

The idea of time scatter plot came up from the practical need to compare the evolution and the relationship of gross domestic product per capita and research and development expenditures of European countries in the same graph. Therefore, our efforts were focused on finding a graphical way to represent simultaneously these two variables and their evolution in time for a subset of European countries.

Traditional scatter plots allow us to represent only two variables (or dimensions) in the same moment, whereas a graph of time series shows the evolution in time for different variables of the same case or only for one variable of different cases. The classical graphing tools are not able to combine in a unique graph the statistical information that we wanted to show. This was the main reason to develop a variant of the traditional scatter plot.

The time scatter plot is an alternative type of graph using Cartesian coordinates to display values across the time for two variables of a set of data. The data is displayed as a collection of points, linked by lines and ending by an arrow, each having the value of one variable determining the position on the horizontal axis and the value of another variable determining the position on the vertical axis. The purpose of the lines and the final arrow is to link the consecutive points in time for each case, resulting in a chain of lines in the shape of an irregular arrow which starts in the first period of time and ends in the last.

The time scatter plot enables a researcher to obtain a visual comparison of two variables in time and helps to determine what kind of relationship there is between the two variables. This approach also allows to integrate in a two axis plot a third additional dimension, the time, providing more information and interpretability to the figure. We present two empirical application of the time scatter plot to demonstrate its usability in practise. The first example demonstrates the already mentioned relationship of R&D expenditures and GDP in Europe for 10 years period of 1995-2005. The second case shows the GDP convergence process of certain European regions.

2 Methodology

The objective of a chart must be to convey the major story being revealed by the data in an unambiguous and illuminating form, transmitting to the researcher or

observer the major quantity of information about a set of data in an interpretable, clear and accurate way.

The basis of our alternative display is the simultaneous representation of two variables for each case in the set of data using Cartesian coordinates and including in the same graph the corresponding points to each one of both time series. Until this point we only have a structure of typical scatter plot with the special characteristic that there are represented points in different moments of time.

Our value added to the graph is that we link the lines between each point of a case in a moment with the next point in time for the same case marking the points in time with arrows. In this way we obtain a chain of linked lines creating an irregular arrow for each case. These arrow shapes start in the first period of the time series and end in the last available period (missing values in the time series can be supported by extrapolation). Therefore, the time scatter plot represents as many chains of arrows as the number of cases of the data. The direction of these arrows shows the evolution in the time of each case for the two selected variables and the position of each point show the relationship between the variables.

Another distinguishing feature of our alternative display is the possibility of divide our chart in four quadrants from the start and/or the end point of a target case. In our approach it was relevant to compare the evolution of certain countries with a concrete target, the European average. Both divisions in quadrants, one related on the initial period and other related on the last period of the objective case, can be also interpreted in terms of speed of convergence or divergence to the target. This is a powerful characteristic of our chart raising the information and improving the interpretability of the chart.

Depending on each variable incorporated to the graph the quadrants can be interpreted in different manners. The context and the domain of study contribute to determinate how the distinct quadrants must be interpreted.

3 Implementation

Several programs provide excellent statistical and graphical interfaces to edit and compose different kind of pre-formatted graphics for compare and analyse information. The problem appears when the idea that you have is impossible to implement with these standard formats or when you need to insert certain additional information to the plot. In all of these cases, open-source software offers a great opportunity to develop and design new alternatives exploiting the code already existing. A clear example of this is Gretl and other statistical software like R-project.

Gretl [1] has implemented several models of graphs to represent the most important kinds of statistical output, in particular, those related to cross section, time series and panel data analysis. Also, with the command line window, the user can add and modify the standard graphs by including additional parameters to the call of the gnuplot function. This function controls the transference of information between Gretl and a plotting utility, Gnuplot, which is used to generate graphs.

Gnuplot [2] is a portable command-line driven interactive data and function plotting utility for different operative systems. The software is copyrighted but freely distributed and support many uses, including web scripting and integration as a plotting engine for third-party applications like Octave, Gretl, etc. This means that any display or graphical model built in Gnuplot could be supported by Gretl. This is a great opportunity to generate new models and graphical designs to explain statistical information, which could be provided by Gretl.

For the implementation of the time scatter plot we have different options. One of them is to download the source code of Gretl and implement the new model of graph by the modification of those functions involved in the plotting process (mainly, in graphing file). An easier option is however to control the flow of information by scripting or launching the command line console of Gretl. Gnuplot function in Gretl accepts literal commands to control the appearance of the plot. In this way it is possible to build the proposed time scatter plot by editing the script with the appropriate commands. These scripts need a lot of numerical information and commands which make them preferable to write by code. Finally, for control in a more detailed way the output, as in this case, it is possible to operate directly, and with the same commands, in Gnuplot application. An easy R-script can generate, as in this case, the literal code for Gnuplot (also the graphical tool of R can support this kind of representation, as in the original version).

4 Practical application

Next two different practical applications of the alternative display are shown, the first of them relative to the R&D effort and its influence in the GDP per capita of the European states and the second relative to one of the traditional objectives of the European Union, the social and economic convergence, analyzing the current situation and the speed of convergence of certain regions of the EU.

4.1 Research and development as driver of innovation and growth

In the context of a global economy and from the Lisbon Strategy point of view, research is a component of a knowledge triangle (the other two being education

and innovation) meant to boost growth and employment in the European Union [3], [4].

Relevance of the case Knowledge has been increasingly acknowledged to be a main driver of economical growth [5], [6]. Consequently, the member states of European Union agreed in 2002 to increase the research and development investments following the Lisbon Strategy. Specifically they set the goal of R&D expenditure to reach 3% of GDP by 2010 [7].

The European Commission intends to play a central role in driving and coordinating European research across the Framework Programmes to have a leverage effect on national research spending, in order to achieve the objective of 3% R&D expenditure of GDP on research in Europe [8].

Analyse the current position and speed of convergence with regard to this target is a relevant purpose for the Commission, comparing the effort of the European nations in R&D expenditures (percentage of GDP) and its impact on the income expressed by GDP per capita.

From a theoretical point of view, it is expected that high level of investments in research and development are resulting in higher level of GDP in future. On the other hand, also the countries with higher initial level of GDP are prone to have relatively higher level of investments in R&D because of more capital available for investing and vice versa the countries with low initial level of GDP are not expected to have very high level of R&D investments.

Showing the results The following time scatter plot shows the R&D expenditures (% of GDP) and GDP per capita (euros) of European Union member countries in years 1995-2005. Each country is having a line with arrows marking each year. The starting point for each arrow is the level of R&D expenditures and GDP per capita in year 1995 (expect for certain countries with missing values in the first periods) and accordingly, the ending point of the arrow remarks the levels of year 2005 (for all the countries).

For a better representation of the points and arrow, we use in this case a square scale in both axes. The black lines crossing the axes are representing the EU-27 averages and the red line crossing the y-axis shows the R&D expenditure target for 2010.

Generally the graph demonstrates rather linear relationship between R&D expenditures and GDP of EU-27 countries -the higher the level of R&D expenditures, the higher the GDP per capita.

Quadrant A shows the observations with higher than EU average spending on R&D but lower than average GDP per capita. As expected, there are very few points in this quadrant. Similarly, there are only view countries with high

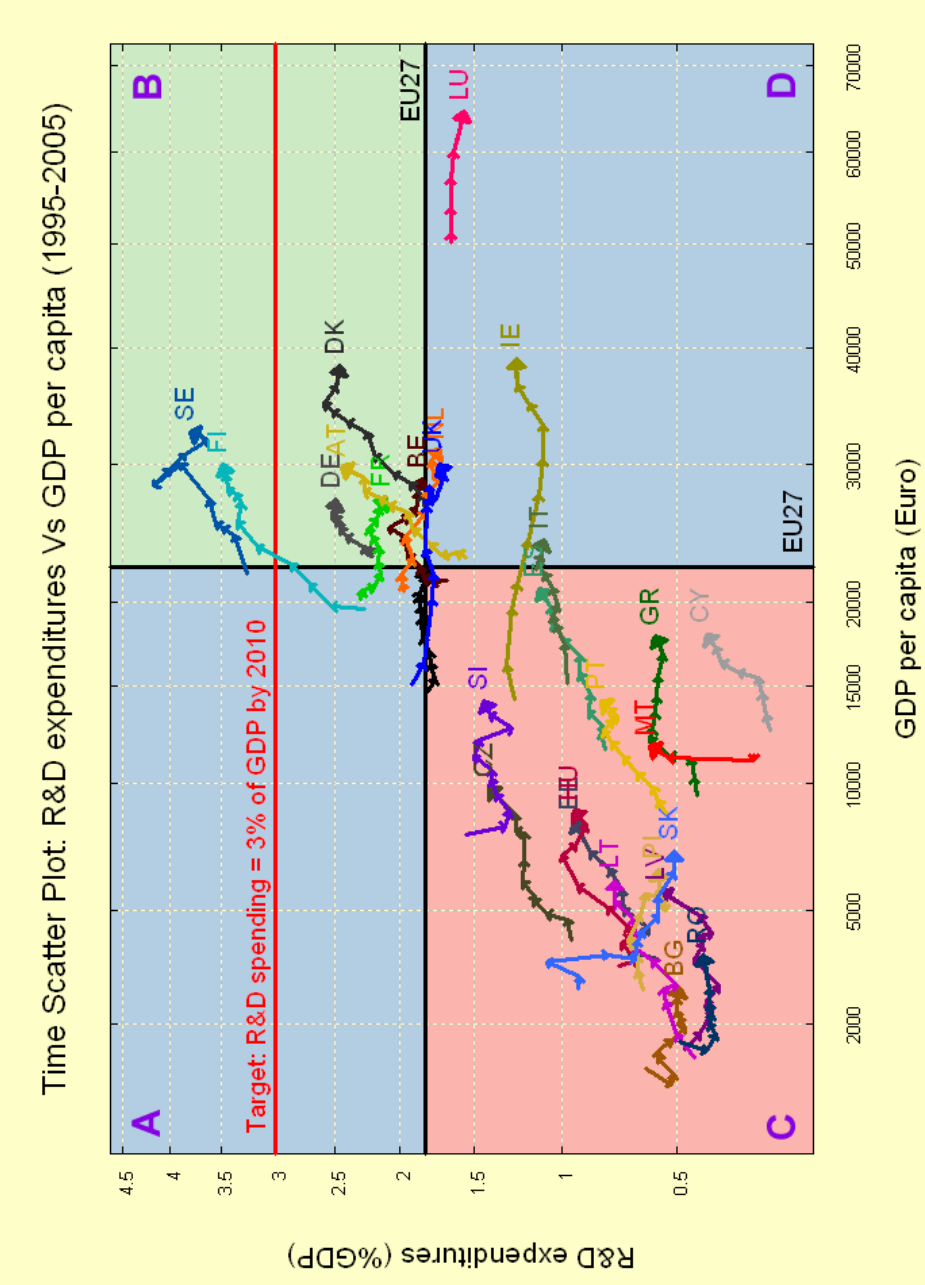


Fig. 1. Time scatter plot: R&D expenditures vs. GDP per capita (1995-2005)

GDP per capita but lower than average spending on R&D as it can be seen from quadrant D.

Most of the EU-15 countries are in quadrant B implying that they have high level of investment on R&D and rather high level of GDP. Only two countries, Sweden and Finland, have reached the EU R&D spending target of 3% but also others like Denmark, Austria and Germany are showing a rather fast growth in R&D expenditures.

The quadrant C represents the countries with below the average R&D spending and GDP per capita lower than EU-27. Relative increases in both R&D expenditure and GDP per capita are the highest in this quadrant.

4.2 The goal of social and economy convergence in Europe

Gross Domestic Product per capita is an excellent indicator for measuring the level and speed of economic and social convergence in Europe. One of the objectives of EU is to increase the welfare of its citizens, reducing the income disparities between different countries and regions.

Relevance of the case Economical convergence has been in interest of economic research for long -first by neo-classical growth theory and later by new growth theory-. Neo-classical growth theory [9] rationalized convergence by diminishing returns on investments of physical capital, thus implying faster growth rates for countries with lower initial level of capital.

The theory expected the economical growth in long-run to be dependent only on exogenous production factors (availability of labour, technological development) and economical convergence between countries occurred when the levels of these factors were equal between economies. However, the majority of empirical evidence was not consistent with this theory as it seemed that richer countries were growing faster than poorer ones.

So called new growth theory on the other hand believed that the growth of economy is endogenous and depends on new knowledge created by the accumulated capital [10], [11]. Instead of diminishing returns, new growth theory suggests that the returns remains steady thus benefiting the countries with higher initial level of capital. The theory also points out how public policies can effect on growth by governmental investments on R&D [12].

European Union's 27 member states form an internal market for 493 million citizens. The economical and social inequalities are still however very large amongst the 268 regions of Europe. Monitoring the convergence or divergence of EU regions is very important as union's ultimate goal is to promote economical, social and territorial cohesion [13].

The recent enlargements of European Union have significantly increased the economical disparities between member states and its regions. The richest member state (measured by per-capita income) Luxembourg, has seven times larger income per capita than Romania. The differences are even larger at regional level where Inner London reaches 290% of EU-27 income per capita average and at the same time the poorest region Nord-Est Romania has only the level of 23% of EU-27 average [14].

To respond this economical disparity of the EU regions, the European Council agreed on the budget of $347 billion on Structural and Cohesion Funds for the period of 2007-2013 [15]. In total, the structural and cohesion funds yield about 30% share of the budget of EU. Majority of these funds (81.5%) are allocated to convergence regions and targeted on investments on infrastructure and human capital. These investments are expected to result on faster growth rate of economy and better employment performance. Consequently, there is a need for monitoring the regional development policies and especially the economical catch up speed of the convergence regions.

Showing the results The Table 1 presents the list of selected regions, all from NUTS 2 level, which we included in the second exercise for showing the applicability of the time scatter plot. In all case, the researcher is the one who must select the number and adequacy of the cases to be represented in the plot.

The criteria for selecting one or other case should answer to the interest or relevance of the study. In our case it was very important to analyze the trend of Basque Country during the last years because of the great effort executed by the Basque governance institutions for reaching the EU15 average. The other regions are shown as reference to our case of interest, facilitating the development of benchmarking activities to identify best practices in other leader regions.

The second time scatter plot (Fig. 2) represents the convergence of GDP per capita for selected EU-15 regions. In this graph, the axes are GDP per capita (euro) and its annual variation rate (%). This figure helps to analyse the convergence (or divergence) of different regions in comparison with the European average.

The European averages of GDP and its growth rates for both the beginning (1996) and ending periods (2005) are marked with black lines crossing the axes. Each line with arrows marks the development of a selected European region, start point of the line showing the level and growth rate of GDP at beginning period (1996) and the arrow ending the line pointing the situation at ending period (2005).

Table 1. Set of regions in the analysis by code (NUTS)

Code	Region Name	Country
DE11	Stuttgart	Germany
DE21	Oberbayern	Germany
DE24	Oberfranken	Germany
DED2	Dresden	Germany
ES21	Basque Country	Spain
FI18	Etelä-Suomi	Finland
GR43	Kriti	Greece
NL11	Groningen	Netherlands
NL31	Utrecht	Netherlands
PT11	Norte	Portugal
PT15	Algarve	Portugal

To explain the second plot we must characterize the different quadrants in the graph, especially, those determinate in the last period of time by European average.

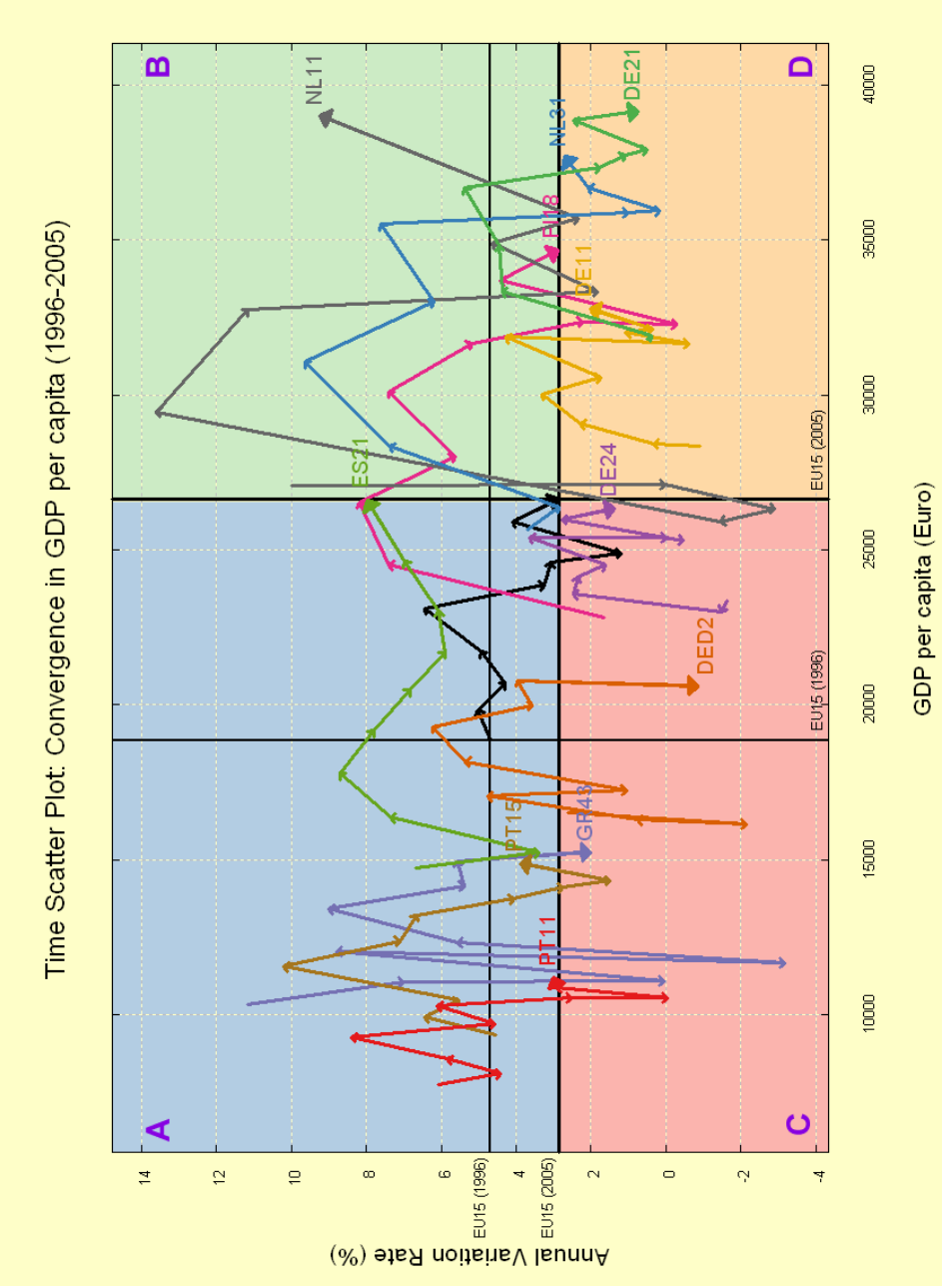


Fig. 2. Time scatter plot: convergence in GDP per capita (1996-2005)

- **Quadrant A (blue colour): Regions with disadvantage but converging**
  This area is relevant to identify regions in a disadvantaged position in terms of GDP per capita, which means these regions have low level of GDP per capita in comparison with the absolute convergence value (the European average). Nevertheless, the regions in this quadrant reached higher annual variation rates. Therefore the relative speed of growth in terms of GDP per capita is higher for all of these regions, generating a very positive process of convergence.

- **Quadrant B (green colour): Regions with advantage and diverging (high relative growth)**
  This quadrant corresponds to regions with a high GDP per capita, higher than the European average. Also these regions have an annual variation rate greater than the mean. Regions in this quadrant are in advanced position, since they have a relatively higher level of GDP and a better variation rate. However, from European point of view, this situation increases the divergence between European regions: rich regions are having a higher growth speed.

- **Quadrant C (red colour): Regions with disadvantage and diverging (low relative growth)**
  In the red area there are placed those regions with disadvantaged position from both point of views. They have low speed of growth and lower level of GDP. In other words, they are regions in disadvantaged economical level and they are diverging instead of converging due lower speed of growth. Consequently, these regions are worsening the European convergence.

- **Quadrant D (orange colour): Regions with advantage but converging**
  Finally, quadrant D covers another converging area. It includes those regions with a good relative level in terms of GDP per capita but with low growth. From convergence view this is a positive situation (Europe progress towards the global convergence), nevertheless these regions must remain alert because of their lower growth rate.

From the point of view of strict convergence in European frame, the optimum position for the regions is in the quadrants A and D. Placing all European regions in both quadrants Europe would reach the perfect convergence in terms of GDP per capita. On the contrary, regions in quadrants B and C contribute to increase the economic divergence in Europe, promoting income inequalities.

Generally, Figure 2 helps us to evaluate, in an analytical way, the condition and the convergence (or divergence) level of different regions in comparison with the European average (EU15) that is considered as reference level.

Analysing the regions by quadrants, Algarve (PT15) and Norte (PT11) regions in Portugal (quadrant A) were initially in unfavourable positions and were growing above the European average during last years. This has reduced the existing gap of these regions with the European reference. Though their position is not favourable yet, the current trend is encouraging.

Similarly, the Basque Country (ES21) starting in the same quadrant with an unfavourable position, has grown well above the European average during last years, reaching in the last analysed year (2005) the European convergence point, maintaining a high speed of growth.

During the observation period the Basque economy presented high rates of GDP growth, which can be explained through some competitive advantages gained from the past -such as appropriate public policies, human capital with high educative level and greater efforts in R&D. In order to maintain the differential growth achieved, the Basque Country may base its future on a clear increment of the productivity level, accelerating the transition towards the economy based on knowledge and technology.

The regions in quadrant B, Etelä-Suomi (FI18) and Groningen (NL11), are in a very favoured position. Their GDP per head was higher than the European average (advantage) and the growth rate was over European Union's average during the whole observation period (divergence trend). Consequently though, they did not contribute towards the aim of converging Europe.

On the other hand, regions like Oberfranken (DE24) and Dresden (DED2) from Germany are placed in the quadrant C. Also Kriti (GR43) in Greece ends up in this quadrant, although the growth rate was more erratic than the average (high variance). The situation for these regions is not so good because they did not reach an acceptable speed of growth and their levels of GDP per capita were not able to reduce the gap with the European average.

Finally, in the last quadrant, the regions of Stuttgart (DE11), Oberbayern (DE21) and, also in the last periods, Utrecht (NL31), began from a very favoured position, but presented a GDP growth below the European Union during the observation period and therefore were converging with the European average. Notwithstanding, the excellent starting positions of these regions have permitted them to keep rather favourable positions also at the end of the observation period.

5 Conclusions

This paper presented a new graphical application to represent time series data called time scatter plot and two empirical cases to illustrate its usability. The advantage of the time scatter plot is that it is able to show the possible correlation

of two variables and in addition their evolvement in time. This is a very illustrative manner to present statistical data, especially with relatively small number of cases.

Flexibility offered by open-source programs like Gretl is essential when applying new modified manners to represent statistical data in graphical forms.

Gretl provides us with an excellent environment to develop and implement new graphical designs and statistical algorithms, due to its flexibility and the accessibility of the source code.

The two empirical cases that we presented as illustrative examples of usability of time scatter plot showed its advantages in presenting time series data of two variables for multiple cases. Although combining the relationship of two variables and dimension of time in same plot may sound complex, the illustration of time scatter plot is very clear and easy to interpret. This is a clear advantage of time scatter plot in the face of end-users of statistical data who may not be experts in statistical models.

In future other type of graphical models applied in different knowledge areas could be adopted to measure and analyze this kind of economic aspects. For example, a violin plot (a combination of a box plot and a kernel density plot) could be applied to show simultaneously the rank and the dispersion of a synthetic index.

Bibliography

[1] Cottrell, A., Lucchetti, R.: Gretl User's Guide - Gnu Regression, Econometrics and Time-series (September, 2008)

[2] Williams, T., Kelley, C.: Gnuplot - An Interactive Plotting Program (March, 2007)

[3] European Council: Presidency Conclusions - Lisbon European Council 23 and 24 (March 2000), http://tinyurl.com/bks2bq

[4] Commission of the European Communities: Facing the challenge - The Lisbon strategy for growth and employment (2004), http://europa.eu.int/comm/lisbon_strategy/index_en.html

[5] Schumpeter, J.: The Theory of Economic Development. Harvard University Press, Cambridge, MA (1934)

[6] Grossman, G. M., Helpman, E.: Innovation and Growth in the Global Economy - Cambridge, MA: MIT Press (1991)

[7] European Council: Presidency Conclusions: Barcelona European Council 15 and 16 (March 2002), http://tinyurl.com/bks2bq

[8] Commission of the European Communities: Implementing the Community Lisbon Programme: A policy framework to strengthen EU manufacturing - towards a more integrated approach for industrial policy. Brussels 5.10.2005. Communication from the Commission, COM (2005)474final

[9] Solow, R.: "A Contribution to the Theory of Economic Growth". Quarterly Journal of Economics 70 (1): 65-94. (1956)

[10] Romer, P. M.: "Increasing Returns and Long Run Growth" Journal of Political Economy, 94, 1002-38 (1986)

[11] Romer, P. M.: "Endogenous Technological Change", Journal of Political Economy 98, 71-102 (1990)

[12] Aghion, P., Howitt, P.: "A Model of Growth Through Creative Destruction", Econometrica 60: 322-351 (1992)

[13] EUR-Lex: Treaty on European Union, Official Journal C 191, (29 July 1992)

[14] European Union: Working for the Regions - EU Regional Policy 2007-2013. (2008) http://tinyurl.com/bqn6q6

[15] Commission of the European Communities: Investing in our future - The European Union's Financial Framework 2007-2013. Second revised edition (March 2008)

Wilkinson Tests and gretl

A. Talha Yalta[1] and A. Yasemin Yalta[2]

[1] TOBB University of Economics and Technology
Sogutozu Caddesi No:43; Sogutozu, 06560, Ankara, Turkey
yalta@etu.edu.tr
[2] Hacettepe University, Turkey

Abstract. Applied econometrics has become fully dependent on computers and software tools. It is therefore important that the reliability of various programs providing econometric functionality is vetted within the profession. Here, we report on the results of our verifying the accuracy of Gretl (GNU Regression, Econometrics and Time-series Library) using the Wilkinson tests. Our study was important in the implementation of a number of modifications improving the general accuracy and reliability of this open source econometric package.

Key words: Gretl, econometric software, accuracy testing, open source

1 Introduction

Like science itself, scientific software is a work in progress and it is possible that any such program at any given time contains errors and imperfections. In the case of econometrics, this is well documented by various authors such as Sawitzki [1], McCullough [2, 3] and Yalta [4], who find important flaws and inconsistencies in the programs widely used within the profession. Because it is nearly impossible to test all of the functionality offered by a typical econometric program, such studies generally employ an introductory or an intermediary test suite such as Wilkinson's 1985 "Statistics Quiz" [5], the Statistical Reference Datasets (StRD) by the U.S. National Institute of Standards and Technology (NIST) or McCullough's set of tests [6]. These procedures are based on comparing the output from a sample of econometric functions against the corresponding correct answers or benchmarked values.

Gretl (GNU Regression, Econometrics and Time-series Library) is a sophisticated and cross-platform program for econometric analysis.[3] It is open source and can be freely used, modified and redistributed under the terms of the GNU General Public License (GPLv3). The program has been gaining in popularity

[3] We first became familiar with Gretl several years ago. Although we never got involved in the coding process of the program, we made contributions in the form of testing the numerical accuracy of its various functions, submitting bug reports, and helping its internationalization efforts.

in the recent years and according to the project's web host SourceForge.net, it was downloaded more than 100,000 times in 2008.[4] See Baiocchi and Distaso [8], Mixon and Smith [9], Yalta and Yalta [10], and Rosenblad [11] for reviews of Gretl versions 0.997, 1.51, 1.6.0, and 1.7.3 respectively.

An important tool offered by Gretl but unavailable in most other statistical packages is the StRD linear regression test suite, which automatically assesses the regression results through a series of 11 tests using the reference data sets compiled by the NIST. This function, which is readily available from the "Tools" menu, helps make sure that a given installation of Gretl produces the certified results, thereby increasing reliability. The so called "Wilkinson Tests" is an alternative entry-level testing procedure also useful for assessing econometric software. Just like the StRD, Wilkinson's method has been widely used for testing different software packages. As a result, the objective of this paper is to describe in detail and report on our experience while applying this procedure, which is currently not available in Gretl.

In the next section, we discuss the Wilkinson tests and their effectiveness in exposing flaws in statistical and econometric programs. In section 3, we report on the various accuracy errors, existence of which we discovered in Gretl versions 1.7.9 and earlier. We also discuss how these errors were fixed following our reporting them to the developers as well as the openness of the whole process, which made it possible to understand the nature of the error and verify its correction directly from the source code. Section 4 concludes.

2 The Wilkinson Tests

The Wilkinson procedure is an entry-level test suite for computational accuracy, which was originally released as a booklet by Wilkinson [5] and described in detail by Sawitzki [12]. The tests are deliberately designed to reveal flaws in statistical software using a small but effective data set NASTY shown in Table 1. As discussed by Wilkinson [13], each column in the table is designed to expose a different type of flaw. For example, ZERO is used for testing the conditions likely to cause various zero divide or singularity errors in computational algorithms. MISS contains all missing values, which are important in some areas of economics. BIG and LITTLE have a significant variation in the eighth digit, making them problematic to analyze using a badly designed program. Together with HUGE and TINY, they are also used for revealing the formatting problems

[4] For the popularity of Gretl, also see the econometric study by Lucchetti [7] (available in this volume) finding that the users of Gretl have been steadily increasing at a yearly rate of 43 percent since 2006.

in various output routines. Finally, ROUND tests how the rounding operation is performed for the purpose of printing.

Although Wilkinson's method is based on an artificial data set, the software defects it is designed to reveal are real as shown by a number of studies such as Sawitzki [1], Bankhofer and Hilbert [14, 15], McCullough [3], and Choi and Kiefer [16]. These studies employ the Wilkinson tests in order to assess the reliability of many statistical and econometric programs, each time exposing deficiencies in fundamental statistical operations such as computing sample standard deviations or graphing. Wilkinson himself argues that the data set is not as extreme as it may seem since, for example, "the values of BIG are less than the U.S. population (and) HUGE has values in the same order as the magnitude of the U.S. deficit." McCullough [17] explains that the procedure has three virtues: First, it is simple and easily applied to most econometric packages. Second, the flaws it is designed to reveal have known solutions so that any program could pass. Third, it questions the functionality that we take for granted such as correctly reading a data file or properly handling the missing values.

The Wilkinson tests are organized in four groups focusing on data management (IA, IB), descriptive statistics (IIA–IIF), missing values (IIIA–IIIC), and linear regression (IVA–IVD) respectively. The first two tests involve reading a custom ASCII data file which includes formatted data likely to be produced by different programs. In the second group of tests, the program first prints ROUND with only one digit. Afterwards, three separate graphs plotting BIG against LITTLE, HUGE against TINY and X against ZERO are produced. This is followed by calculating various summary statistics as well as a correlation matrix and Spearman correlations on all the variables. None of these computations should imply a problem for a well designed program. In Test IIE, X is tabulated against X using BIG as a case weight. This is a strictly statistical procedure not available in most econometric programs including Gretl. Finally, Test IIF involves regressing BIG on X in order to check whether the correct answer BIG=99999990+1X is returned. Test IIIA and IIIB assess the handling of the missing values by running the operations "`IF MISS=3 THEN TEST=1 ELSE TEST=0`" and "`IF MISS=<missing> THEN MISS=MISS+1`" respectively. The answer is 2s or missing values for the first case and all missing values for the second case. Similarly, Test IIIC tabulates MISS against ZERO. The program should return one cell with 9 cases in it. The fourth group of tests first extends the data set by the powers $X1 = X^1, \ldots, X9 = X^9$ and runs a series of four regressions. In Test IVA, X is regressed on X1 through X9. Here, R^2 should be unity since this is a perfect fit. Test IVB regresses X on X and a constant with the obvious solution X=0+1X. This is followed by a regression of X on BIG and LITTLE to test whether the program will warn about the singularity

Table 1. The Data Set NASTY

X	ZERO	MISS	BIG	LITTLE	HUGE	TINY	ROUND
1	0	NA	99999991	0.99999991	1e+012	1e-012	0.5
2	0	NA	99999992	0.99999992	2e+012	2e-012	1.5
3	0	NA	99999993	0.99999993	3e+012	3e-012	2.5
4	0	NA	99999994	0.99999994	4e+012	4e-012	3.5
5	0	NA	99999995	0.99999995	5e+012	5e-012	4.5
6	0	NA	99999996	0.99999996	6e+012	6e-012	5.5
7	0	NA	99999997	0.99999997	7e+012	7e-012	6.5
8	0	NA	99999998	0.99999998	8e+012	8e-012	7.5
9	0	NA	99999999	0.99999999	9e+012	9e-012	8.5

problem. Finally, ZERO is regressed on a constant and X, with the expectation of a warning about ZERO having no variance or a regression output where both the correlation and total sum of squares are given as 0.

3 The Performance of Gretl

During our testing of Gretl, we found a number of errors, which mainly affect the display of the data and the presentation of various computation results.[5] The first problem that we found was in Test IB and was regarding the lack of a formatting function for a "display data" window showing more than one variables. The lack of this functionality has the potential to mislead the users. For example, the default format Gretl uses is printing such values with 6 significant digits. In the case of the Wilkinson data set, this leads to an output wrongly implying that LITTLE is constant for observations 1 to 5, while BIG is constant through 5 to 9. We reported this problem to Gretl developers and within three days there was an update adding the "reformat" function to all data display windows.

The second issue that we encountered was in Test IIA and was due to the fact that the Gretl commands `print`, `printf`, and `sprintf` rounded numerical values using "unbiased rounding" or the "round-to-even" method. As a result, attempting to print ROUND with only one digit returned {0, 2, 2, 4, 4, 6, 6, 8, 8} instead of the correct answer of printing the numbers from 1 to 9. This problem was because of not following the principle that rounding for the purpose of printing is not about calculation but about the presentation of the results. Gretl developers acknowledged this issue and corrected it within just four days.

Thirdly, in Test IIB, Gretl failed to plot BIG against LITTLE accurately and also refused to plot X against ZERO returning an error message. These were

[5] This report is abbreviated. See Yalta [18] for a more detailed discussion.

due to a number of communication problems with the Gnuplot program, which handle the plotting functionality of Gretl. Within six days of our reporting of the errors, the Gretl source files handling the graphing behavior were revised. As can be seen in Figure 1 below, Gretl now correctly shows a 45 degree line and a vertical line for the two plots.

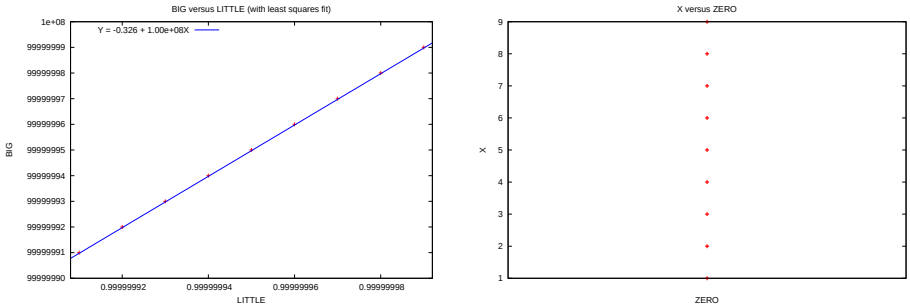


Fig. 1. Gretl's "BIG versus LITTLE" and "X versus ZERO" Plots After Corrections

A fourth error that we came across was while attempting to compute the various descriptive statistics for the eight series in Test IIC. Gretl performed these calculations accurately, however, printed out after rounding the standard deviation of TINY as effectively 0 instead of the correct value 2.7386e+012. This error was also fixed within 24 hours of our reporting it.

Finally, we discovered in Test IID that Gretl produced an erratic output for the calculation of Spearman's rank correlations between ZERO and the other variables. This problem, which was due to numerically printing the special "Not Available DouBLe" (NADBL) value was corrected by the Gretl developers in a matter of 24 hours as well. Gretl passed the remaining 10 tests successfully by computing the results accurately and showing the correct behavior as required by these tests.

In addition to a number of worthwhile improvements, an important benefit of our testing Gretl using the Wilkinson tests was being able to observe how the access to the programming code made it possible to see the cause, facilitated a rapid fix and enabled the verification of the various corrections. Table 2 shows some revision details on the various Gretl source files updated within just a few days after our exposing of the various software defects. It is possible to examine all these changes using SourceForge's "viewvc" interface available at `http://gretl.cvs.sourceforge.net/viewvc/gretl/`.

Table 2. Some Revision Details on the Updated Gretl Source Files

TEST	SOURCE FILE	REV.	DATE	TEST	SOURCE FILE	REV.	DATE
IB	lib/src/printout.c	1.375	Dec 20	IIA	lib/src/printscan.c	1.24	Dec 21
IB	gui2/series view.c	1.44	Dec 18	IIA	lib/src/printscan.c	1.25	Dec 22
IB	gui2/series view.c	1.45	Dec 19	IIB	lib/src/graphing.c	1.409	Dec 19
IB	gui2/series view.c	1.46	Dec 19	IIB	lib/src/graphing.c	1.410	Dec 23
IB	gui2/series view.c	1.47	Dec 20	IIB	lib/src/graphing.c	1.412	Dec 23
IB	gui2/series view.c	1.48	Dec 21	IIB	lib/src/graphing.c	1.413	Dec 23
IB	gui2/series view.c	1.49	Dec 22	IIB	lib/src/graphing.c	1.414	Dec 24
IIA	lib/src/printout.c	1.372	Dec 19	IIB	lib/src/graphing.c	1.415	Dec 24
IIA	lib/src/printout.c	1.373	Dec 19	IIB	lib/src/graphing.c	1.416	Dec 24
IIA	lib/src/printout.c	1.375	Dec 20	IIB	lib/src/plotspec.c	1.42	Dec 24
IIA	lib/src/printout.c	1.376	Dec 22	IIB	lib/src/plotspec.c	1.43	Dec 24
IIA	lib/src/printout.c	1.377	Dec 22	IIB	lib/src/gretl_matrix.c	1.393	Dec 23
IIA	lib/src/printout.c	1.378	Dec 22	IIC	lib/src/gretl_matrix.c	1.388	Dec 19
IIA	lib/src/printscan.c	1.21	Dec 19	IID	lib/src/gretl_matrix.c	1.392	Dec 23
IIA	lib/src/printscan.c	1.22	Dec 19	IID	lib/src/graphing.c	1.411	Dec 23
IIA	lib/src/printscan.c	1.23	Dec 20				

4 Final Thoughts

It is not extraordinary or uncommon that a complex econometric program such as Gretl contains errors and imperfections. The important issue is the mechanism through which such problems are addressed by the developers. Earlier studies on the reliability of econometric packages show that software vendors are unequal in their attention to computational accuracy. Sawitzki [1] ran the Wilkinson tests on nine different commercial packages, found a number of errors in all them and reported that the reaction he received from different vendors varied ranging between “cooperative concern and rude comments.” Yalta [4] found that the various numerical issues in the GAUSS software package reported by Knusel [19, 20] and later by Vinod [21] were not fully fixed in seven years and after several major revisions. Yalta and Jenal [22] report that Addinsoft did not fix the grossly erroneous least squares estimator for ARIMA in the XLSTAT statistical program and let the users obtain invalid results by using a defective function. Microsoft Excel is widely used in the field of economics and McCullough and Heiser [23] discuss that errors found in Excel97 were still either not fixed or wrongly fixed in Excel2007. On the other hand, there also exist studies such as Zeileis and Kleiber [24], Keeling and Pavur [25] and McKenzie and Takaoka [26] which report correction of errors or more accurate results in comparison to earlier versions of various econometric packages.

A question worth investigating is whether the transparent and collaborative nature of the open source development model provides some advantages in the process of error correction, resulting in better and more reliable software in a scientific setting. Indeed, McCullough [27] finds that the open source Gnumeric spreadsheet program fixed within a few weeks all the reported flaws surprisingly similar to those found in Excel. Similarly, Kuan [28] examines three pairs of commercial and open source programs and reports generally faster fixing of bugs in the latter. Our experience applying the Wilkinson tests on Gretl was concurrent with these cursory studies. We observed the correction of all the reported flaws after just six days. This is a remarkable performance considering the fact that the developers do not receive any monetary compensation for their contributions to the program. In addition, here it was also possible for us to access the source code and see the exact cause of the problem each time we discovered an error. Furthermore, the open source nature of Gretl also enabled an instant dissemination of the various fixes and enabled our verifying the correction of the errors.

In conclusion, our assessment of Gretl using the Wilkinson tests allowed the detection of several important flaws and resulted in a number of worthwhile revisions. Also, it is our understanding that the availability of the programming code and the absence of commercial concerns can provide an open source econometrics package such as Gretl an advantage in the reliability department.

Acknowledgments

We wish to thank Allin Cottrell for providing prompt and useful answers for our reports and inquiries regarding Gretl and its inner workings.

Bibliography

[1] Sawitzki, G.: Report on the numerical reliability of data analysis systems. Computational Statistics and Data Analysis **18** (1994) 289–301

[2] McCullough, B.D.: Assessing the reliability of statistical software: Part II. American Statistician **53** (1999) 149–159

[3] McCullough, B.D.: Wilkinson's tests and econometric software. Journal of Economic and Social Measurement **29** (2004) 261–270

[4] Yalta, A.T.: The numerical reliability of gauss 8.0. The American Statistician **61** (2007) 262–268

[5] Wilkinson, L.: Statistics Quiz. 1 edn. SYSTAT, Evanston, IL (1985)

[6] McCullough, B.D.: Assessing the reliability of statistical software: Part I. American Statistician **52** (1998) 358–366

[7] Lucchetti, R.: Who uses gretl? an analysis of the SourceForge download data. In: Proceedings of the 1st Gretl Conference, Bilbao, Spain (Forthcoming)

[8] Baiocchi, G., Distaso, W.: GRETL: Econometric software for the GNU generation. Journal of Applied Econometrics **18** (2003) 105–110

[9] Mixon, J.W., Smith, R.J.: Teaching undergraduate econometrics with GRETL. Journal of Applied Econometrics **21** (2006) 1103–1107

[10] Yalta, A.T., Yalta, A.Y.: GRETL 1.6.0 and its numerical accuracy. Journal of Applied Econometrics **22** (2007) 849–854

[11] Rosenblad, A.: Gretl 1.7.3. Journal of Statistical Software **25** (2008) 19

[12] Sawitzki, G.: Testing numerical reliability of data analysis systems. Computational Statistics and Data Analysis **18** (1994) 269–286

[13] Wilkinson, L.: Practical guidelines for testing statistical software. In Dirschedl, P., Ostermann, R., eds.: Computational Statistics, 25th Conference on Statistical Computing at Schloss Reisenburg, Physica, Verlag (1994)

[14] Bankhofer, U., Hilbert, A.: Statistical software packages for windows - a market survey. Statistical Papers **38** (1997) 377–471

[15] Bankhofer, U., Hilbert, A.: An Application of Two-Mode Classification to Analyze the Statistical Software Market. In: Klar, R., Opitz, O., eds.: Classification and Knowledge Organisation. Springer, Heidelberg (1997) 567–572

[16] Choi, H.S., Kiefer, N.M.: Software evaluation: EasyReg international. International Journal of Forecasting **21** (2005) 609–616

[17] McCullough, B.D.: The Accuracy of Econometric Software. In: Belsley, Kontoghiorghes eds.: Handbook of Computational Econometrics. Wiley (to appear)
[18] Yalta, A.T.: Should economists use open source software for doing research? (2009) [Unpublished Manuscript].
[19] Knüsel, L.: On the accuracy of the statistical distributions in GAUSS. Computational Statistics and Data Analysis **20** (1995) 699–702
[20] Knüsel, L.: Telegrams. Computational Statistics and Data Analysis **21** (1996) 116
[21] Vinod, H.D.: Review of GAUSS for Windows, including its numerical accuracy. Journal of Applied Econometrics **15** (2000) 211–220
[22] Yalta, A.T., Jenal, O.: On the importance of verifying forecasting results. International Journal of Forecasting **25** (2009) forthcoming
[23] McCullough, B.D., Heiser, D.A.: On the accuracy of statistical procedures in Microsoft Excel 2007. Computational Statistics and Data Analysis **52** (2008) forthcoming
[24] Zeileis, A., Kleiber, C.: Validating multiple structural change models – a case study. Journal of Applied Econometrics **20** (2005) 685–690
[25] Keeling, K.B., Pavur, R.J.: A comparative study of the reliability of nine statistical software packages. Computational Statistics and Data Analysis **51** (2007) 3811 – 3831
[26] McKenzie, C.R., Takaoka, S.: Eviews 5.1. Journal of Applied Econometrics **22** (2007) 1145–1152
[27] McCullough, B.D.: Fixing statistical errors in spreadsheet software: The cases of Gnumeric and Excel (2004) [CSDA Statistical Software Newsletter; retrieved December 10, 2008].
[28] Kuan, J.: Open source software as consumer integration into production (2001) [Unpublished Working Paper, Stanford University].

Subject index

Author index