Newsletter of the Political Methodology Section

American Political Science Association
Volume 6, Number 1, Fall 1994

Co-Editors: R. Michael Alvarez, California Institute of Technology
Nathaniel Beck, University of California, San Diego

Contents

Notes From the Co-Editors

R. Michael Alvarez
California Institute of Technology

Nathaniel Beck
University of California, San Diego

This issue is devoted largely to articles about the various software packages political methodologists use. Unfortunately, we do not cover every software package available to political methodologists in this issue — in future issues we will cover others. Here we have articles which span a number of different software packages, including GAUSS, S-PLUS, LIMDEP, EViews, MicroTSP, and RATS.

We also have several contributions on the issue of replicability, including the text of the "Statement on Statistical Reporting, Archiving, and Replication." This was developed by a committee composed of prominent political methodologists, and was presented at the section business meeting during the APSA meetings in New York. There it was debated and approved by the section. We include here both the text and a discussion from Walter Mebane (committee member). Please read both carefully, and forward your comments and ideas to the members of the committee, Larry Bartels (Section President), and journal editors.

Of related interest, Gary King discusses archives for the storage and distribution of replication datasets, and announces the new editorial policy of *The Political Methodologist* to publish brief citations of the availability of these datasets. In a similar spirit, Jim Stimson here provides an updated "policy mood" dataset. (*The Political Methodologist* will, in general, only publish announcements relating to the archiving of data. Only in rare cases will we publish hard copy of that data.)

Finally, we present the second part of "Exogeneity, Inference and Granger Causality" by Jim Granato and Renee Smith. There are several book reviews in this issue, and section news.

While this issue of *The Political Methodologist* focuses on econometric software packages and initiatives to promote reporting and replication in political science research, the next issue of *The Political Methodologist* will be primarily concerned with teaching of political methodology. We welcome all submissions on this subject — from syllabi to more general discussions of the issues associated with teaching political methodology at any level. We will obviously not be able to print many syllabi, but we hope to be able to announce an ftp site for archiving syllabi.

The Data Analysis Revolution and S-PLUS

Christopher H. Achen
University of Michigan

Successful statistical packages make conveniently available to researchers the statistical tools found in textbooks of the prior decade.[1] When the first mainframe packages like SPSS appeared in the 1970's, they offered contingency tables, factor analysis, and some other basic techniques. But their most common use for publication was to carry out linear regression estimated by ordinary least squares (OLS).

In the familiar regression notation, if y_i is the ith observation on the dependent (endogenous) variable and X_i is the corresponding vector for the independent (exogenous) variables, then:

$$y_i = X_i\beta + u_i \quad (i = 1, \ldots, n) \tag{1}$$

where u_i is an unobserved disturbance (error) term and β is a coefficient vector to be estimated. As is well known, under the usual Gauss-Markov or normal linear model assumptions, OLS estimation of the coefficient vector is statistically optimal under a variety of prominent theoretical criteria. Analysis of variance, essentially another way of doing regression, has similar theoretical attractions and was included in the early packages for the same reasons.

The econometric tradition that held sway until the 1980's saw linear regression as foundational. Textbooks began with such a model and then proceeded to complicate it. The goal was to reach the generality of relationships such as:

$$y_i = f(z_i, u_i; \theta) \tag{2}$$

where now z_i is a vector of explanatory variables and the function $f(.)$ is assumed known up to a vector of parameters θ, which are to be estimated. Thus if y_i and $f(.)$ are vectors, one has "seemingly unrelated regressions," and if the vector y_i is included in z_i, then "structural equations." If not all the elements of z_i are observed without error, one faces errors in variables or unobserved variables. If the disturbances are not additive or are not white noise, one seeks to transform or modify the equation (2) until they are. If $f(.)$ is neither linear nor transformable to linearity, one carries out nonlinear estimation. And so on. Successive waves of mainframe packages, beginning with SAS and including TSP and specialized programs like LISREL, embodied this philosophy.

In the framework of that period, the task of econometric theory was to estimate the parameter vector "efficiently," that is, with minimum mean squared error (MSE) among some class of unbiased or consistent estimators. Thus much effort was expended by practitioners to show that they had corrected for heteroskedasticity or for serial correlation, because it was known that, asymptotically at least, these estimators produced lower MSE. How much lower, or whether MSE really was lower in finite samples, were questions avoided in practice, since the relevant theory was usually messy or absent. In any case, the real scientific evaluation of the project did not depend on which statistical correction was done. (Or if it did, the entire research enterprise was called into doubt.) Rather the point was to silence the narrower kind of methodological critics by performing up to the contemporary standard of what econometric theory understood well.

But what did theory understand well? The answer is that it knew how to answer questions of this kind: Suppose in equation (2) that we are perfectly certain of the functional form of $f(.)$ and have no doubts about the density of u_i (or at least know its lower moments). Then theory could tell us how we might best estimate θ. Method of moments or minimum chi-square estimators did not differ in this respect from maximum likelihood estimators (MLE). All assumed that functional forms relating explained variables to explanatory variables were known with certainty up to a finite parameter vector. In general, MLE and Bayesian approaches required certitude about the distribution of the explained variable as well. In all cases, much was made of how well the estimators reduced the uncertainty that was left over, namely that concerning the coefficients.

In practice, of course, uncertainty about coefficients is a relatively small fraction of our ignorance. We often have only hunches about the effects of individual variables, and no clue as to how they combine to influence the explained variable. Sociological or psychological theory is usually completely useless in this respect, and even when good economic theory is available, it is often silent about the function $f(.)$. Thus functional forms are unknown, and the distribution of the unobserved variables composing the disturbance is more theological still. However, the available econometric theory forced us to pretend that all these mysteries were thoroughly understood in advance.

[1] Thanks to Larry Bartels, Neal Beck, Gary King, and especially Simon Jackman for their comments and assistance. The views expressed are my own.

This way of thinking about social science specifications has had pernicious consequences. How often have we seen students and even colleagues innocently estimating regression models with a dozen or more explanatory variables? Most of the time, all the regressors are assumed without argument to enter the model linearly. If some minor econometric difficulty is present (a little heteroskedasticity or first-order serial correlation, or a dichotomous dependent variable), how often have we seen these same researchers choosing to spend their limited time switching to a more trendy estimator which they could not readily interpret, or laboriously "correcting" their procedure to produce nominal efficiency, only to end with estimates which differed from their OLS counterparts by just a fraction of their standard errors?

In these same studies, basic data analysis is often omitted because no time is left for it or because the techniques have never been learned. Unfortunately, elementary partial regression plots and leverage point analysis all too often demonstrate that the specification errors due to nonlinearities and omitted variables are two or three orders of magnitude more important to the true MSE than the econometric fixes. But in the contemporary climate, it has often been better to estimate a dopey linear model with heteroskedasticity and serial correlation corrected than to estimate a more nearly correct nonlinear model with OLS. After all, researchers remember, getting up to 100% efficiency in the linear case is what their textbooks and instructors taught them.

Thus linear models persist. Even in fields like voting behavior, where the classic nonlinearities constitute the empirical foundation of the subject, linear regressions and probits predominate. Researchers meet their morning classes and tell the sophomores how voters' defection from party ID is a curvilinear function of political knowledge: Voters at intermediate levels of information are more likely to defect than either the more ignorant or the more sophisticated. Then in the afternoon, they return to their offices and instruct their statistical package that vote choice is a fixed linear function of party ID and political knowledge.

Crude linear regressions have their uses, of course, and every experienced researcher has run some of them. But the style has come under increasingly heavy fire in political science, following the pattern in economics a decade ago. Why believe models which depend so heavily on dramatically implausible linearity assumptions, especially when no effort is made to adduce evidence in their favor? Why not just stop reading at that point and move on to the next article?

Among scientifically oriented researchers, the low credibility of most empirical social science specifications may induce either of two reactions. The first is to demand stronger mathematical theory to structure statistical work. Macroeconomists have had some luck using control-theoretic ideas as starting points for deductions ending in statistical specifications. They have developed more general estimation methods to cope with the resulting nonlinearities and cross-equation restrictions, and created statistical tests to assess the assumed functional forms. Some of the best recent texts, such as Davidson and MacKinnon (1993), provide statistical support for this approach. They spend far less time on the linear case and far more on nonlinear models than was typical a decade ago. They also emphasize specification tests and robust standard errors.

Time series methods are also of growing importance, since the credibility of cross-sectional work is so low. Hamilton's (1994) book, appropriate for a post-regression course, downgrades structural equation estimation to brief, skeptical treatment, for example, while stochastic difference equations, the Kalman filter, vector autoregression, and cointegration get the kind of extended treatment many have come to feel they deserve. Work of this kind calls on econometric packages for personal computers and work stations such as `GAUSS`, `RATS`, `SST`, and, especially for the control-theory aspects, `MATLAB`, `MATRIX-X`, and to some degree `GAUSS` as well.

The alternative response to the crisis in econometrics is more agnostic and more radical. Instead of replacing the god of linearity by the goddess of rationality, this school of thought believes only what it can demonstrate for itself. Specification error gets very heavy emphasis; other econometric problems are treated as second-order. Plotting and other visual data-analytic techniques get priority; interactive desktop computing is essential. High-powered estimation occurs only at the end when the researcher can make a strong *empirical* case for the necessary specification assumptions. This approach has been popular in statistics and in sociology, neither of which has much tradition of using rigorous causal theory to structure empirical work. But clearly, the same techniques may also be used to complement statistical models derived from game theory or other formal theory; there is no reason why data analysis must always be done barefoot.

A sizable share of the inspiration for this approach to statistics derives from John Tukey. Some middle-aged readers may remember the pleasures of Mosteller and Tukey (1977), the first textbook I know that took regression analysis seriously as both science and art. Its successors have been legion, particularly in the last decade. Tools proliferate daily for finding patterns in data, and no regression course is complete without attention to them. The best-known ideas are incorporated into good intermediate-level statistical packages of recent years, such as `SYSTAT`. But as Simon Jackman notes elsewhere in this issue of *TPM*, no package takes the new approach as seriously as does `S-PLUS`, and no other package provides anything like the range and power of `S-PLUS` for doing modern data analysis.

S-PLUS was originally written to run under UNIX, and until recently most users have accessed it on multi-user minicomputers or have installed it on their own UNIX workstations. However, a Windows version has become available, making the package accessible to a broader group of researchers. In this review, I will focus on S-Plus for Windows, version 3.1, which I've been using since the first of the year.

The package installs easily. All the plotting and statistical options I have tried behave as advertised, and the program has never hung or bombed in ten months of exploration and production use. In short, the port from UNIX seems very successful.

S-PLUS is designed for serious data analysts, and it is more demanding to use than, say, RATS. (However, GAUSS's hold on first place for cantankerousness is not in danger.) A few features of the S-PLUS package require some getting used to. Its object oriented programming requires that if one wishes to list the functions or datasets one has created (the "objects"), one must type `objects()`. (The initially more natural command `objects`, without parentheses, produces a listing of the code for carrying out the command.) This style is typical of S-PLUS: Even exiting the package requires parentheses. But after awhile, the logic of this approach comes to seem natural.

Other irritations are more long-lasting. Chambers and Hastie (1992) demonstrate that many of the statistical procedures in the parent program S have a common structure and can be carried out by similar commands. Nevertheless, much remains to be done to make the user's life easier. Even among the various plotting and regression-type estimators, the order of entering variables (dependent variable first or last? Separated by comma or tilde? Or do both work?) varies haphazardly, as do the commands for displaying results.

The particular implementation of S-Plus for Windows adds a few minor problems of its own. The newly released factor analysis procedure contains only the estimation procedures of the 1950's. It has no option for the modern method of doing MLE with identifying constraints on the loadings (as in Lawley and Maxwell, 1963). Some of the graphics management procedures are a little clumsy, too. For example, if one requests a plot, the command fails unless one has already opened a graphics window with the unmemorable command `win.graph()`. Once the window is open, the plot works, but the display does not switch to that window; one has to click on the window to see it. In fact, the only time the display *does* switch automatically to the graphics window is at its initial opening–at which time the window is blank, of course.

Perhaps the largest single hurdle for new users of S-Plus for Windows, however, is the current state of the user manuals supplied with it. While some sections are capably written, others are not, and a few exhibit a capriciousness, opacity, and disorganization that will inevitably remind academic users of sophomore essays. In addition, the index for the user's manual is so foolishly constructed that one cannot help laughing at its frequent lapses into uselessness. Fortunately, most of the crucial topics are treated in Chambers and Hastie (1992), where the writing is a model of exposition. Reading it as needed sidesteps most of the problems, especially in conjunction with the on-line help in the program. Use of the S-PLUS manuals may then be reserved for learning the specific details of running the program under Windows.

In the end, most of these difficulties with S-PLUS are minor and soon overcome. The important news is that S-PLUS is in a class by itself for data-analytic power, and that overall, working with it is a joy. First, the graphics displays are excellent. For example, with a click of the mouse, plotted points may be labeled with the name of the observation (such as "Thomas Jefferson" or "Utah"), making it easy to give substantive interpretation to outliers and high-leverage points in regression analysis. The graphics are also easy to control, label, and print. The same work that lets the researcher understand the data also produces photocopy-ready displays for oral and written presentations. Data cleaning and coding are remarkably easy, too–an unusual feature in high-level packages.

Second, the large collection of conventional and robust estimation techniques are well integrated with the graphics. Histograms and cross-tabulations are easy to use and attractively displayed. In regression, GLIM, time series, or analysis of variance contexts, one can interactively display the data along with various conventional, robust, or nonparametric versions of the fit. Best of all, the list of statistical procedures in S-PLUS is not static. Recent monographs often list the S-PLUS functions for computing their estimators (e.g., Hardle, 1991), and in other cases, the relevant code has already been incorporated into the package itself (e.g., McCullagh and Nelder, 1989, and Efron and Tibshirani, 1993). S-PLUS is sufficiently rich in statistical functions that programming estimators on one's own is often relatively easy; for instance, the many attractive extensions to probit and logit set out in Morgan (1992) deserve implementation in S-PLUS functions and will no doubt come into wider use over the next few years. Libraries of S programs are available on the Internet as well, although it should be noted that many of these are written for UNIX and require some retouching to run under Windows.

S-PLUS also provides valuable assistance with nonlinear estimation procedures, such as those used in most MLE and some Bayesian applications. For my taste, other packages too often encourage a frontier mentality about nonlinear estimation: one writes down the objective function, tells the program to compute approximate numerical derivatives, and then shouts, "Let 'er rip." One can do the same in S-PLUS, too, but better alternatives are provided conveniently. In common with numerical optimization texts at every level, Chambers and Hastie's (1992) guide to S takes a critical attitude toward numerical derivatives and computed standard errors. Hence S-PLUS provides help in profiling the likelihood function to assess how well an optimization routine has performed on a given likelihood. Strongly globally concave objective functions maximized with analytical derivatives may be an exception, but most nonlinear estimations shouldn't be trusted without this kind of testing. (See also Brady, 1990.)

Above all, with S-PLUS one does a lot more checking and takes a lot less for granted, simply because the package makes it easy to do so. And with looking, one finds things. Of course, most readers of this newsletter will be familiar with at least the basics of graphical display, outlier detection and leverage analysis in regression, and will have used some robust methods. But having a very wide range of the newest and most powerful tools of that kind conveniently available in a single statistical package makes more difference than one might expect.

I have been surprised several times at how much I learned from using S-PLUS to plot various kinds of robust fits against a dataset and its subsets, even with data I thought I knew well and whose plots I had examined visually many times. For example, the nonparametric "smoothers" often find kinks and dips in relationships. Most of these seem to be sampling error, but some are thought-provoking, and an occasional one arrives as a blinding revelation. In the latter cases, the usual OLS, method of moments, and MLE techniques would have brutalized the data.

Even robust regression fits, while less agnostic than smoothers, occasionally raise one's eyebrows. In a figurative sense, they estimate relationships in the manner of a case study analyst: they are willing to down-weight or even drop whole chunks of observations to display the most common causal sequence. Thus I have seen a highly robust fit reverse the sign of a bivariate OLS regression slope in a small-to-medium-size dataset consisting of coded historical events. Sampling error is always a potential cause, particularly since sampling theory for many robust methods is not well understood. However, in the instance I encountered, deletion of substantively odd observations and those with doubtful codings essentially allowed OLS to reproduce the robust fit—a strong signal that more than one statistical regime was present in the data and that the usual methods would have been seriously misleading. Without checking, how would one have known?

For the vast majority of empirical researchers, of course, it is much too soon for a wholesale conversion to robust and nonparametric estimation. Those techniques are in a stage of rapid development, and a sorting out will have to occur, along with a consolidation of the underlying theory, before most social scientists will understand them well enough to make sensible use of them and leave older methods behind. Thus least squares ideas, which date to the nineteenth century, and MLE, which was developed in the 1930's, will remain the workhorses of daily production. S-PLUS has most of their standard applications available in convenient form. At least for the present, though, the package contains relatively few of the usual econometric estimators for structural equation systems, which depend so heavily on strong assumptions about the data. The latter methods undoubtedly contain substantial specification errors in actual applications, but there are times when ignoring selection bias or simultaneity imposes even larger errors. Thus owning S-PLUS is not a replacement for having packages like SST, GAUSS, or RATS around.

Nevertheless, after a few months of using a package like S-PLUS, work habits begin to change—and attitudes, too. One begins to wonder how many of the fits reported in the journals, from the most routine to the most innovative and computer intensive, would survive data-analytic checking of the kind that S-PLUS encourages. For the estimates typical of the literature, with blithely assumed linear functions of exogenous variables, no checking by robust or nonparametric techniques, no displays of partial regression plots, no profiling of likelihood functions, and no diagnostics beyond t-tests, R^2, and log-likelihoods, credibility drops to near zero, no matter how state-of-the-art the econometric techniques or how powerful the theorizing.

To a growing number of observers, the older estimation methods, carried out in conventional fashion, have begun to look like the tools of the torturer, cattle prods applied to the data's tender flesh. Fewer and fewer believe the confessions those techniques have extracted. The new statistical approaches of the last decade have aimed at letting the data speak the truth without fear. At the moment, S-PLUS comes closest to letting us hear what they have to say.

References.

Brady, Henry. 1990. Review of Gill, Murray, and Wright's *Practical Optimization. The Political Methodologist.* 3(1): 5–8.

Chambers, John M., and Trevor J. Hastie. eds. 1992. *Statistical Models in* S. Pacific Grove, California: Wadsworth and Brooks/Cole.

Davidson, Russell, and James G. MacKinnon. 1993. *Estimation and Inference in Econometrics.* New York: Oxford University Press.

Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap.* London: Chapman and Hall.

Hamilton, James D. 1994. *Time Series Analysis.* Princeton, New Jersey: Princeton University Press.

Hardle, Wolfgang. 1991. *Smoothing Techniques with Implementation in S.* New York: Springer-Verlag.

Hastie, Trevor J., and Robert J. Tibshirani. 1990. *Generalized Additive Models.* 2nd ed. London: Chapman and Hall.

Lawley, D. N., and A. E. Maxwell. 1963. *Factor Analysis as a Statistical Method.* London: Butterworths.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models.* 2nd edition. London: Chapman and Hall.

Morgan, B. J. T. 1992. *Analysis of Quantal Response Data.* London: Chapman and Hall.

Mosteller, Frederick, and John M. Tukey. 1977. *Data Analysis and Regression.* Reading, Massachusetts: Addison-Wesley.

Review of LIMDEP 7 (beta)

Nathaniel Beck
University of California, San Diego

I have been teaching a two week course at the Essex Summer School for a number of years. This course, similar to the maximum likelihood course at the ICPSR Summer Program, stresses limited dependent variables and event history analysis. I have been using GAUSS, making primary use of programs written for Gary King's Count module combined with some programs of my own. While the flexibility of GAUSS is wonderful, GAUSS is not the easiest program to use and I am not very good at writing general packages. As a result, more time was spent explaining GAUSS and/or my particular programs than was spent discussing statistical issues. (Since Essex has only the Unix version of GAUSS I was unable to use King's very nice integrated Count package.) Hence I went in search of a new package.

SAS does everything, and hence must include all that I need. But it is a huge program, and I have never found it easy to either use or teach with. Other general packages (such as SPSS or BMDP) don't do all that I need or are less widespread and as clunky as SAS. I looked for a more tailored program, one that a student could purchase in a PC version but one that would also run in a UNIX environment. There are many excellent programs (Shazam, SST) that handle the standard econometric models very well. Both also have the facility to estimate additional maximum likelihood models. But neither of these programs is particularly strong on event history analysis. This process of elimination led me to LIMDEP. LIMDEP, written by William Greene, the author of *Econometric Analysis*, has been around for almost a decade. It is currently available in Version 6, though I have used a beta copy of Version 7. Version 7 should be available Spring, 1995. It runs both under DOS and UNIX; the DOS version works well OS2.

LIMDEP has always contained an incredible set of limited dependent variable routines, both for discrete and count data. These have been augmented in Version 7 with a full nested multinomial logit module (allowing up to three choice levels and a maximum of 70 choices) estimated by full information maximum likelihood. LIMDEP is the only general program I know that can do this. (Now if it could only do multiple probit!) The new version also contains a series of panel routines for limited dependent variables. It is also quite good at selection bias, and has some useful things such as double probit, as well has having a very nice non-parametric binary choice model based on Manski's work.

LIMDEP also has an excellent set of routines for doing parametric duration modelling, as well as for doing Cox proportional hazard modelling. Of great importance to me, it can handle time varying covariates and discrete duration models. It also handles more specialized problems, such as split-sample duration models. In short, it seems to contain most everything needed for the analysis of duration data.

Like the other good econometric packages LIMDEP has all the standard econometric routines, as well as a built-in matrix language and the facility to estimate general maximum likelihood models. LIMDEP is clearly not going to be as fast as GAUSS, but it is nice to have the additional flexibility when needed. (LIMDEP has so many built-in routines that it is unclear how often one would want to write a special purpose maximum likelihood program in LIMDEP.) The matrix language is of great use in allowing for a variety of tests; the manual is very good at pointing out how those tests might be done.

Many people have found the user interface of LIMDEP difficult to get used to. Version 7 has improved the interface, and, it is promised, will offer considerably more on-line help. (The written manual is excellent, both in describing how to use the program and also in discussing various econometric extensions.) There are now a series of F-key driven menus, giving the user easy access to a file-manager, editor and status indicator. While LIMDEP 7 makes primitive use of a mouse, it is far from a windowed program. I use it under OS2, which allows me to edit programs in emacs while keeping the LIMDEP window open. I find it more difficult to use LIMDEP in DOS, but others may be able to remember variable names for more than a second. (In any event, LIMDEP is no worse than any other non-windowed program.)

LIMDEP on the PC produces VGA quality graphics which can be dumped to a laser printer. The graphics are adequate. Hard core Bell Labs types will never confuse LIMDEP with S-PLUS, but it's graphics are on a par with other DOS based econometric software.

In short, LIMDEP is a full-featured econometric program that contains a number of more specialized procedures that are ideally suited for the types of data that we often see in Political Science and Sociology. When a colleague came to me with a problem that turned out to be a double-probit, I initially had a moral dilemma: do I offer to write the double-probit routine in GAUSS or not? Then I remembered LIMDEP and my conscience was clear. My colleagues, my students, and I are all happier when I can leave the program writing to Bill Greene. I still use GAUSS when I need raw speed for simulations or when I need to program Kalman filter routines, but I find that much of my research and almost all of my teaching can be better done in LIMDEP.

LIMDEP is available from Econometric Software, 43 Maple Ave., Bellport, NY, 11713. Phone: 516-286-7049; FAX: 516-286-6848. Price approximately $600. Version 6 available now, Version 7 in beta test, available early next year. Runs under MSDOS (8Mb RAM desirable) and UNIX.

EViews: MicroTSP for Windows and the Macintosh

Harold Clarke
University of North Texas

The econometric software bazaar is a crowded one with venerable "command-line" oriented programs like RATS, SST and TSP jostling for market share with menu-driven ones such as MFIT (formerly DFIT) and PCGIVE. A new entry in the latter category is EViews (Econometric Views). EViews is the Windows version of Quantitative Micro Software's well-known MicroTSP, which has been available since 1981. Why would (should) you be interested in EViews? To paraphrase the now famous sign, "it's the GRAPHICS stupid!" Although visual displays are crucial for the description and analysis of time series data, the graphics capacities of econometrics programs have lagged far behind their statistical power. Recent improvements notwithstanding, most popular programs still do not do graphics very well. EViews is a welcome exception. Single and multiple graphs can be edited (extensively) and moved about the screen at will, and there are several printing options. Moreover, since EViews is a Windows program, you can easily import graphs into documents.

EViews' superior graphics are accompanied by an extensive menu of basic and advanced statistical routines. These include, inter alia, garden-variety OLS regression (with AR and MA parameters or PDLs if desired), two- and three-stage least squares, nonlinear least squares, probit and logit routines for binary choice models, and VAR (not BVAR) with impulse response and variance decomposition. Forecasting capability for single-and multiple- equation models is included as well. Like two of its competitors (MFIT and PCGIVE), EViews has a suite of model diagnostics to test for autocorrelation, heteroskedasticity (general and ARCH), normality, functional form and parameter constancy. It also does Dickey-Fuller unit-root tests (with MacKinnon's critical values), and Johansen tests for cointegration. Again, many of these tests are accompanied by razor-sharp graphics that can be converted into "objects" to be saved for later study, printed, imported into documents, or transferred other EViews workfiles.

Since it is menu-driven, EViews is very easy to use, and you (and your students) can concentrate on the modeling rather than on trying to recall the syntax (or arcane error messages when you get it wrong) of a command-line program. Like most of its competitors, EViews reads ASCII data, standard spreadsheet formats and, if you wish, you can enter data directly from the keyboard. Not surprisingly given its lineage, EViews also reads MicroTSP data sets. Data transformations are easy (all standard operators and functions are available), and working with subsamples of your data is no problem.

What kind of hardware do you need? The requirements are not trivial – the minimum is a 386 PC, four megs of RAM, five megs of free hard disk space and, of course, Windows 3.1 and a rodent. Since it is a Windows program, heavy-duty hardware helps. Although I have not clocked it against its rivals, EViews is very fast. Running on a Gateway 60 MHz Pentium with 16 megs of RAM and an ATI 32-bit graphics board, analyses of multivariate time series models with 154 observations and several right-hand side variables seem instantaneous – click the mouse, and the results are there. Joining the "frequent regressors club" (apologies to SHAZAM) was never easier.

Is EViews, then, the answer to a time series analyst's number-crunching dreams? The answer is "yes," if you are content with the reasonably large (but hardly exhaustive) suite of statistical routines provided. However, if you want to write your own specialized routines or use those written by others, EViews may not do everything you want. Unlike its cousins (e.g., MFIT, PCGIVE), EViews has a programming language, but you still may want to have other programs, e.g. RATS or GAUSS, for some tasks. RATS will shortly release a Windows version (its in BETA as this is written) and, if the Estima folks come close to getting the graphics right, they will reduce the big comparative advantage EViews currently holds in this area. Even then, however, the ease of use of programs like EViews will continue to make them attractive for many research and teaching purposes. But, why choose just one; let's challenge the scarcity assumption and have our departments buy whatever combination of programs suits our tastes. No problem, right?

EViews, by David M. Lilien et al. is available (for $295) from Quantitative Micro Software, 4521 Campus Drive, Suite 336, Irvine, CA 92715. Phone: 714-856-3368; FAX: 714-856-2044

GAUSS and S+: A Comparison

Simon Jackman
University of Chicago

Comparing statistical software for an audience of quantitative political scientists is asking for trouble.[1] To express a preference for one piece of software is to risk scorn from otherwise collegial fellow social scientists. People who agree on most issues at department meetings can suddenly find themselves at each other's throats should the perennial SAS vs. SPSS-X question arise. We (sometimes rightly) perceive high costs in learning the ropes of a second or third package, and migrating "pet" data sets from package to package. I suspect most political scientists using computers to analyze data are firmly committed to one software package, and may occasionally dabble with a "specialty" package; examples include LISREL or EQS for analyzing covariance structures, and TSP or RATS for time series. Sometimes choices over hardware or operating systems, perhaps imposed by budget-sensitive deans and department chairs, make one piece of software more attractive than another. Given these constraints, general purpose packages such as SAS or SPSS-X, running on large multi–user machines, probably remain the most commonly encountered statistical computing environments in political science.

In this essay I compare two pieces of software not as well-known as SAS or SPSS-X, but well-suited to the needs of political methodologists, and "work–a–day" empirical social scientists more generally: GAUSS and S-PLUS. GAUSS is more widely used and familiar to political scientists than S-PLUS, and so I will provide more details about S-PLUS than GAUSS. My aim here is to compare the two packages, rather than review each in turn.

I should make own preferences clear from the start. While fairly easy–going on software issues I find S-PLUS a superior package for my "day–to–day" statistical computing, and I regularly use GAUSS for computer–intensive applications. Further, by focusing on GAUSS and S-PLUS I am not suggesting that these two software packages are the be–all and end–all in statistical computing. While both GAUSS and S-PLUS are getting better on the data–*management* front (as distinct from data–analysis), I use SAS for getting data into reasonable shape for analysis. And neither package provides as much support for estimating structural time–series models or covariance structures as one would like.

Above all, I want to stress the virtue of a pluralist approach to software choice. Software is nothing more than a tool for tasks we have to perform. For a given task, some tools will be better than others. But there is more at stake here. As I show below, both GAUSS and S embody different ideas about analyzing data. It is important to keep these different ideas in mind when using these packages on a day–to–day basis. There is a real danger of starting to view data analysis in terms of what a favorite piece of software can do, or does well. The two pieces of software I consider here have strengths that could conceivably cause a narrowing of methodological vision; perhaps the diehard user of GAUSS risks reducing every problem to an exercise in maximum likelihood estimation, while the dedicated S user might agonize endlessly over functional–form, robust estimators, outlying and influential data points, and so on. Reasonable people might disagree as to which of these is the worse vice. But once methodological problems start being perceived or even defined in terms of what one's favorite software does well, then the software has stopped being a tool, and has become at best a crutch, and at worse a shackle. The pluralist stance I advocate here follows a more general pluralism I hold regarding methods and techniques for empirical social science. But down to basics.

GAU::

GAUSS is the package of choice for "high-end" quantitative analysis in political science. *The Political Methodologist* contained a "GAUSS Corner" under previous editors, and prominent political methodologists have sung the praises of GAUSS in published work (e.g., Beck 1989, 121). The popularity of GAUSS for sophisticated quantitative work stems in part from the growing abundance of maximum likelihood estimators in the work of methodological specialists. For instance, King's work on event count models (1988, 1989a) and Beck's (1989) exposition of the Kalman filter both rely heavily on GAUSS's maxlik routine, Beck going so far as to claim that without GAUSS and maxlik he could not have conducted his analysis at all. Furthermore, King's (1989b) more general advocacy of maximum likelihood is of practical value only to the extent that political scientists have access to appropriate software. GAUSS's maxlik is almost without peer in this regard. Given an arbitrary, user–defined likelihood function, maxlik will return estimates of the parameters maximizing the likelihood and the variance-covariance matrix of the parameters, without user-supplied routines for the calculation of first derivatives or a Hessian matrix, though these can be supplied if known and

[1] Chris Achen, Mike Alvarez, Larry Bartels, Neal Beck, Charles Franklin, David Gow, Don Green, Gary King, Walter Mebane, Gillian Weiss, and Bruce Western helped me develop, think, and write about my own software preferences, but are not responsible for what follows.

programmed. "Do–it–yourself" MLE is a reality for political scientists largely because of GAUSS and **maxlik**.[2]

GAUSS has several other strong–points for quantitative political scientists. First, GAUSS is extremely fast. This is particularly useful given the computational burdens of MLE, derivative-free or not. Further, most quantitative political scientists lack formal training in computer programming, but GAUSS runs fast enough so as to often make program-efficiency a trivial concern. I am always surprised by how quickly GAUSS can crunch through an inelegant routine for evaluating a likelihood function during a **maxlik** run, or repeatedly loop over a hastily written, inefficient set of instructions for, say, bootstrapping, Monte Carlo simulation, or EM-type approaches to estimation (e.g., Gelman and King 1990; Jackman 1994). For this reason GAUSS is my preferred package for most of my computer–intensive applications.

Second, GAUSS works with matrices directly, and the syntax for manipulating matrices in GAUSS closely follows the conventions of matrix algebra. Given X (an n by k matrix) and y (a n by 1 vector) the coefficients of the regression of y on X can be obtained in GAUSS simply by entering `inv(x'x)*x'y`, though calling the GAUSS function `ols` is simpler: `call ols(0,y,x)`. For this reason GAUSS makes an excellent teaching tool. Textbook formulae can be entered almost directly (as above), allowing students to "learn by doing."[3]

:-P –U:

S–PLUS is an ensemble of programs for statistical analysis using the S language. S was developed at AT&T and almost in its very structure reflects the orientation of Bell Labs statisticians such as John Tukey, William Cleveland, and (more recently) Trevor Hastie. Reflecting the philosophy of those statisticians, S, and its' friendlier superset S–PLUS, are oriented towards "data visualization." In a windowing environment S–PLUS presents the user with graphics and help windows (I use OpenWindows or X on a UNIX machine directly, or remotely via a PC running as an X server). Users can interact dynamically with their data via graphics windows; for instance, one can "brush" and "spin" data clouds, or "point and click" with a mouse to identify outlying observations. S–PLUS also contains *many* robust regression estimators and smoothers[4] all of which are tailored to present their results graphically. GAUSS seems a long way behind S–PLUS on this front. GAUSS comes with no robust regression estimators and its graphics output lags far behind S–PLUS's in quality, flexibility, and ease–of–use.

The emphasis on data–visualization in S–PLUS becomes more apparent when one contrasts it against GAUSS's strength in MLE.[5] MLE starts with a set of assumptions about the stochastic and structural form of a model. **maxlik** returns a set of estimates after a researcher (a) assumes (consciously or not) the likelihood function completely characterizes the data, and (b) programs a routine to evaluate the likelihood function given the data and a set of parameter values. S–PLUS permits and encourages a different approach. The emphasis on data visualization orients the S–PLUS user towards thinking through what assumptions about the stochastic and structural parts of the model might be appropriate. For instance, is a linear fit an adequate summary of the structural relationship between y and x? If not, then the choice of stochastic assumptions used to write out a likelihood function (normal errors versus poisson, log–normal, Cauchy, or whatever) may well be moot. But if the structural assumptions are reasonable, then in many cases the choice of stochastic assumptions may well be less crucial (e.g., the normal is a reasonable approximation for several widely–used distributions). All this is to say that S–PLUS allows researchers to easily "look at the data" and, if nothing else, get a feel for how well the data conform to the structure they might want to impose in a particular ML setup. GAUSS, while providing quick and easy MLE, lacks comparable data visualization, and may well be "putting the cart before the horse," so to speak.

Second, S is a very open and flexible "object–oriented" language. Generic functions in S first determine what kind of object is being passed to the function, which then call the method appropriate to that class of object. Two commonly used generic functions include `summary()` and `plot()`, which I discuss briefly below. S–PLUS recognizes many classes of objects including vectors, matrices, data frames ("data sets"), the results of fitting a regression or some other statistical model, functions, a formula to be used in a statistical model, a list of cut–points with which to categorize data, a family of distributions to be used as link and variance functions in fitting a generalized linear model (GLM, see below), or lists (collections of objects). Users can create new classes of objects and write appropriate methods for generic functions.

To give some idea of the openness and power of the object–oriented approach, consider an object `x`, a numeric vector. The command `summary(x)` will print the mean,

[2] Of course, derivative-free MLE via numeric optimization is not without its drawbacks. A discussion of those issues is beyond the scope of this short essay. A useful discussion with reference to GAUSS's **maxlik**, is Henry Brady's (1990) review of Gill, Murray and Wright (1981). See also Green (1991).

[3] The introductory version of Judge et al.'s (1988) classic econometric text has a companion GAUSS computer handbook (Hill 1989).

[4] e.g., Cleveland's et al. (1992) `loess()` and Hastie and Tibshrani's (1990) `gam()` functions. See Chambers and Hastie (1992).

[5] Bruce Western helped me develop the argument in this paragraph; the standard disclaimer applies.

median, and some quantile information on `x`. If instead `x` is an object produced by fitting a linear model then the `summary(x)` command would have instead printed a summary of the linear model (parameter estimates, standard errors, some summary statistics on the residuals, an estimate of the residual variance, the correlation matrix of the parameter estimates). In turn the results of `summary()` could be assigned to an object.[6]

GAUSS embodies some features of the object–oriented approach, but there are basically only three classes of GAUSS "objects": matrices (and vectors, both numeric and character), GAUSS data sets, and GAUSS functions. Almost all GAUSS functions take matrices as their arguments and return matrices.[7] S on the other hand is far more general in its use of objects, and huge dividends can result. Consider the following S-PLUS command:

```
reg1 <- lm(approval ~ elect.cycle + unemp +
   infl + sp500)
```

This command creates the object `reg1` which contains the results of fitting a linear model to `approval` using the regressors `elect.cycle`, `unemp`, `infl`, and `sp500`. Now consider the following:

```
reg2 <- update(reg1, ~ . - sp500,
       subset=(year != 195C))
```

The `update()` function creates a new object, `reg2`, by dropping the `sp500` variable from the regression, and dropping the observation for 1956. Both `lm` objects are now available for comparison via calls to the appropriate functions. For instance, `anova(reg1,reg2)` will produce an F test comparing the two specifications.

Other "generic" functions in S-PLUS include the `plot()` function. Methods for `plot()` exist for many classes of objects, consistent with the "data visualization" philosophy. Many different objects can be passed to `plot()`, and the action taken by S-PLUS will depend on the particular attributes of the object passed. Passing a `lm` or `glm` object (created with the linear model or generalized linear model functions, respectively) will return plots of the residuals against the dependent variables, and prompt the user for a series of partial residual plots.

Comparing Strengths and Weaknesses

How to compare S-PLUS and GAUSS? Consider the strengths of GAUSS: an excellent optimization routine well–tailored for maximum likelihood, speed, and "direct" programming of matrix algebra. S-PLUS offers superb "interactivity" via graphics and a flexible object-oriented approach. But how does S-PLUS measure up on GAUSS's strengths?

While S-PLUS doesn't have anything as general or as quick as `maxlik`, it does have three excellent optimizing functions, `ms()`, `nlmin()` and `nlminb()`; all perform derivative–free optimization of user-supplied functions, while the latter allows the user to specify constraints on parameters. These functions are implementations of public–domain optimization routines (Gay 1983) and are fast and fairly reliable. The only drawback is that without user–supplied derivatives the routines will not provide an estimate of a Hessian matrix upon convergence, and hence, standard errors of MLEs are not returned as a matter of course as they are with GAUSS's `maxlik`. This is a real drawback for non–standard ML problems where derivatives are costly to compute, program or even derive in the first place.

But all is not lost. For problems that can be represented as non–linear least squares the S-PLUS function `nls()` will provide derivative–free estimates of parameters and standard errors (see Bates and Chambers (1992) for examples). S-PLUS also contains a remarkable `glm` function with which many "seemingly-ML" problems can be estimated. Many limited dependent variable setups (e.g., logit, probit, and poisson regression) can be written in the McCullagh and Nelder (1989) general linear models (GLM) framework and easily analyzed in S-PLUS. GLMs generalize the standard linear model

$$y = X\beta + \epsilon, \qquad (1)$$

by positing two functions:

- a *link* function, $g(\cdot)$, that describes how the mean of $y \equiv E(y) = \mu$ depends on the linear predictors: $g(\mu) = X\beta$
- a *variance* function that captures how the variance of y depends upon the mean: var(y)=$\phi V(\mu)$, with ϕ constant.[8]

A wide class of models can be written in this framework. For instance, a logistic regression of voter turnout on a set of regressors (e.g., Nagler 1991, 1994) can be computed in S-PLUS with the command:

```
turnout.logit <- glm(turnout ~ education +
     education^2 + age +
     age^2 + south +
     gub.elec + closing.date +
     I(closing.date*education) +
     I(closing.date*education^2),
     family=binomial)
```

where the defaults accompanying the `family=binomial` option select the logistic regression model from the class

[6] One especially useful function is `latex()`, which dumps the contents of an object into a form suitable for reading by LaTeX.

[7] An exception is the `lreg` function which returns regression results in a single stacked vector, which can then be examined by passing the stacked vector to an extracting function. This is a small step towards the object–oriented approach.

[8] See Hastie and Pregibon (1992, 196–7).

of GLMs: a logit link, $g(\mu) = \log(\mu/(1-\mu))$, and a binomial variance function, $V(\mu) = \mu(1-\mu)/n$. Many families of links and variance functions are available, including the `gaussian()`, `poisson()`, `Gamma()`, and `inverse.gaussian()`. Different links may be specified with these families, including probit and complementary log-log links in addition to the logit example given above. The `quasi()` family provides a way for users to specify arbitrary link/variance combinations, while the `power()` link function provides a way for users to parameterize a link.[9]

The GAUSS approach to these problems is to have the user write a function to calculate the log-likelihood, and let `maxlik` go to work. But for the wide class of problems expressible as a GLM S-PLUS allows the user to easily try alternative specifications of functional form etc without a re–write of the function GAUSS's `maxlik` requires for the log-likelihood. Further, all the generic S-PLUS functions with methods for `glm.objects` are available (e.g., `plot()`, `anova()`, `summary()`, `update()`, `predict()`).

So, there is a case to be made that one can live without GAUSS's `maxlik` if one is creative enough with the `glm()` and `nls()` functions in S-PLUS. But what about GAUSS 's other great strength, speed?

There is no question S-PLUS loses the speed race to GAUSS. There are all sorts of reasons for this. GAUSS compiles itself and user instructions into "binary pseudo-code" which is then executed deep and quickly in a computer's CPU. GAUSS is available for shared UNIX systems, but most users run it on stand–alone 486 PC's. Currently, S is not a compiled language and while PC versions are available, most users run it on shared UNIX platforms, subject to the vagaries of processor and network load. My limited experience with the Windows version of S-PLUS suggests that it is perhaps even slower than the UNIX version. Looping and iterative tasks run slowly in S, though "vectorizing" (or even "matricizing") can speed up things considerably. And because S is a more general language than GAUSS, it is less straightforward in programming matrix manipulations: GAUSS's elegant `inv(x'x)*x'y` becomes `solve(t(x)%*%x)%*%t(x)%*%y` in S.[10]

Perhaps the real question is when ought we care about speed? Non–standard ML problems, Monte–Carlo simulations, EM, or Gibbs samplers are seldom encountered on a daily basis, even for methodologists. Though, when these tasks are undertaken, it is nice to have results sooner rather than later. GAUSS's speed with these tasks remains the primary reason I continue to use it. Otherwise I find S-PLUS's data visualization and object-oriented approach a more attractive environment for my day–to–day statistical computing.

Getting acquainted with the openness and flexibility of the object–oriented approach of S-PLUS involves some startup costs. I was introduced to statistical computing on IBM mainframes running SAS jobs in batch, and I then migrated to the more interactive GAUSS. Finding my way around S-PLUS took some time, as did GAUSS. But S-PLUS comes with extensive on–line help, organized into menus, and accessible with a mouse. Entries in the S-PLUS help menus typically conclude with references to the relevant statistics literature and examples using data sets supplied with the software. After trying out some of these examples I became increasingly proficient with the language and able to exploit the emphasis on data visualization and the object–oriented approach.

In addition, a key feature of any software package is the ability of the user to make extensions to the product. Both S-PLUS and GAUSS are easily supplemented with user–written functions, but I find S-PLUS slightly superior to GAUSS in this regard. User–written S-PLUS functions require none of the annoying declarations that accompany GAUSS functions, and simple debugging of user-written functions is provided as the functions are edited from within S-PLUS. Arguments can be passed to S-PLUS functions in a very open and flexible way, allowing users to write much more general types of functions than GAUSS permits.

I raise two points in closing, both to do with the "culture" of both packages and their user groups. First, S is constantly growing and is probably the dominant computing environment in applied statistics. New techniques are often created in S and deposited in the S archive at `lib.stat.cmu.edu` (also known as StatLib to many gopher or WWW servers, etc). The S-news mailing list frequently contains contributions from eminent statisticians announcing new S functions available for downloading from StatLib, or providing help with using existing S functions. A GAUSS mailing list also exists (to which I admittedly don't belong), but I know of no GAUSS archive as comprehensive and as widely–known as StatLib. Gary King's server at `haavelmo.harvard.edu` contains many `maxlik` examples and is a great asset for the GAUSS user community, but is a far cry from StatLib.

The role of these public electronic repositories should not be underestimated. Functions that first appear on StatLib later become part of the S-PLUS ensemble, and the authors are typically cited in work using the functions. Publicity, often in the way of citations, and a desire to contribute to the S-PLUS user–community typically motivates StatLib contributors. GAUSS, on the other hand, grows through proprietary third party add–ons, which often lag behind

[9] Political scientists' work in this direction typically employs direct MLE or non–linear least squares. Viewed from a GLM perspective, Nagler's (1994) "scobit" estimator is essentially a parametrization of a link function for a logit or probit model. Also, King's generalizations of event count models (1989a) have a straightforward interpretation in the GLM framework.

[10] This example for the OLS estimator of β is perhaps specious, since both GAUSS and S-PLUS have more efficient functions for calculating solutions to frequently encountered problems like this (e.g., forming crossproduct matrices, etc).

cutting–edge techniques, since authors are obliged to get routines into a highly polished state for commercial distribution. The informality of StatLib, coupled with a strong ethos of coöperation among S-PLUS users means that a good proportion of the "plus" part of S-PLUS remains free and in the public–domain. A public–domain GAUSS server will help to generate the coöperation that exists among S-PLUS users. This server will also give users access to many more GAUSS functions, which until recently, have been written from scratch by individual users, e–mailed from user to user, or (some years later) purchased as a third–party add-on.

Second, it is no accident that GAUSS is preferred by econometricians and S by statisticians. S is more in the data–analytic tradition of statisticians like Tukey and Mosteller than it is geared toward "high–brow" econometrics. Conversely for GAUSS. One of my more recent mailings from GAUSS's makers, Aptech, carried news of Fair-Taylor and DynGames, two modules for solving rational expectations models, available for $80 each.[11] Contrast some recent news on S–news; Bill Cleveland reminded S users that all the techniques from his book *Visualizing Data* will be commercially available as a S-PLUS library this summer, and John Chambers reported on developments underway to improve S.

Both S and GAUSS go some way towards embodying distinct methodologies. S is fairly explicit about its orientation toward a data–driven, data–analytic methodology. Compared to S, GAUSS appears to come more from an assumption–driven, model–testing approach. Either package could benefit from incorporating some of the other's strengths. But given that empirical political science appears to employ a mix of both approaches, or at least ought to, both GAUSS and S-PLUS are tools worth having around. More specifically, if the task at hand involves substantial looping or iteration, and speed is a factor, then GAUSS is hard to beat. But for the rest of my statistical computing, in which I value flexibility and graphical displays of my data and results, S-PLUS is my tool of choice.

GAUSS is available from Aptech Systems, Inc., 23804 S.E. Kent–Kangley Rd, Maple Valley, WA 98038. ph: (206) 432–7855. fax: (206) 432–7832. bbs: (206) 432–8377 ($50/year). A user–group can be joined by sending a request to gaussians-request@uclink.berkeley.edu

[11] Other third party add-ons for GAUSS include LINCS, a linear covariance structure analysis package; MISS ($115 for the pair, and both written by maxlik's author, Ron Schoenberg), a program for imputing missing data; and COINT ($80), a library of procedures for evaluating and estimating cointegrated systems.

S-PLUS is available from StatSci (a division of MathSoft, Inc), 1700 Westlake Ave. N., Suite 500, Seattle, WA 98109. sales dept: (800) 569-0123. ph: (206) 283–8802. fax: (206) 283–8691. e–mail: mktg@statsci.com The S-news user–group can be joined by sending e-mail to s-news-request@utstat.toronto.edu

Table 1: Selected price information as of April 1994 (academic users only).

GAUSS-3.1	PC	UNIX
single user	$495	$595
network copy	$745	$595[a]
Network users:		
2–5 users	$325 per	$325 per
6–10 users	$150 per	$150 per
11+ users	$125 per	$125 per
Descriptive Stats	$75	$155
Quantal Response	$75	$155
Maximum Likelihood	$95	$185
Linear Regression	$75	$155
Time Series	$95	$185
Splus 3.2		
single user[b]	$887	$1380

[a] UNIX prices vary depending on platform, and are subject to a yearly maintenance and support fee after first year. Third party applications available for PC only. Discounts available for purchases of three or more application modules.

[b] UNIX prices may vary depending on number of simultaneous sessions permitted. PC version requires Microsoft Windows. Maintenance and support fee after first year applies to both versions.

References.

Bates, Douglas M., and John M. Chambers. 1992. Nonlinear Models. In Chambers, John M. and Trevor J. Hastie. eds. *Statistical Models in* S. Wadsworth and Brooks/Cole. Pacific Grove, California. 421–54.

Beck, Nathaniel. 1989. Estimating Dynamic Models Using Kalman Filtering. *Political Analysis.* 1:121–156.

Brady, Henry. 1990. Review of Gill, Murray, and Wright's *Practical Optimization. The Political Methodologist.* 3(1):5–8.

Chambers, John M., and Trevor J. Hastie. eds. 1992. *Statistical Models in* S. Wadsworth and Brooks/Cole. Pacific Grove, California.

Cleveland, William S., Eric Grosse, and William M. Shyu. 1992. Local Regression Models. In Chambers, John M. and Trevor J. Hastie. eds. *Statistical Models in* S.

Wadsworth and Brooks/Cole. Pacific Grove, California. 309–76.
Gay, D.M. 1983. Algorithm 611: Subroutines for Unconstrained Minimization Using a Model/Trust–Region Approach. *ACM Transaction in Mathematical Software.* 9:503–24.
Gelman, Andrew, and Gary King. 1990. Estimating the Consequences of Electoral Redistricting. *Journal of the American Statistical Association.* 274–82.
Gill, Philip E., Walter Murray and Margaret H. Wright. 1981. *Practical Optimization.* Academic Press. London.
Green, Donald. 1991. Maximum Likelihood for the Masses. *The Political Methodologist.* 4(2):5–11.
Hastie, Trevor J., and Daryl Pregibon. 1992. Generalized Linear Models. In Chambers, John M. and Trevor J. Hastie. eds. *Statistical Models in* S. Wadsworth and Brooks/Cole. Pacific Grove, California. 195–248.
Hastie, Trevor J., and Robert J. Tibshirani. 1990. *Generalized Additive Models.* Chapman and Hall. London.
Hill, R. Carter. 1989. *Learning Econometrics Using* GAUSS*: A Computer Handbook to Accompany Judge et al.* 2nd edition. Wiley. New York.
Jackman, Simon. 1994. Measuring Electoral Bias: Australia, 1949–93. *British Journal of Political Science.* 24:319–57.
Judge, George G., R. Carter Hill, William E. Griffiths, Helmut Lutkepohl and Tsoung-Chao Lee. 1988. *Introduction to the Theory and Practice of Econometrics.* 2nd ed. Wiley. New York.
King, Gary. 1988. Statistical Models for Political Science Event Counts: Bias in Conventional Procedures and Evidence for the Exponential Regression Model. *American Journal of Political Science.* 32:838–63.
King, Gary. 1989a. Variance Specification in Event Count Models: From Restrictive Assumptions to a Generalized Estimator. *American Journal of Political Science.* 33:762–84.
King, Gary. 1989b. *Unifying Political Methodology.* Cambridge University Press. Cambridge.
McCullagh, P., and J.A. Nelder. 1989. *Generalized Linear Models.* 2nd edition. Chapman and Hall. London.
Nagler, Jonathan. 1991. The Effect of Registration Laws and Education on U.S. Voter Turnout. *American Political Science Review.* 85:1393–1405.
Nagler, Jonathan. 1994. Scobit: An Alternative Estimator to Logit and Probit. *American Journal of Political Science.* 38:230–55.

Review of MicroTSP and RATS

Renée M. Smith
University of Rochester

Given the rapid pace of changes in econometric theory for time series data, a frequently-heard question among analysts is: "What statistical package do you/did you use to estimate your results?" In what follows, I discuss the advantages and disadvantages of two software packages – MicroTSP v. 7.0 and RATS v. 4.10 – that I use to complement each other when I am analyzing time series data.

MicroTSP, by Quantitative Micro Software, was developed solely for use on personal computers. With pull-down menus and numerous interactive prompts, MicroTSP is a user-friendly program that easily can be learned and used by students in a first-year regression course. Chapter Two of the User's Manual walks new users through examples of creating, manipulating, and displaying data and of estimating an OLS regression. Unlike other software programs with complicated documentation, when I loan a MicroTSP manual to a student, I generally do not see that student again until s/he is *finished* with the manual.

MicroTSP can be installed on IBM-compatible personal computers where it requires up to 1.2 mb of disk storage. The program is issued on both 5-1/4 inch and 3-1/2 inch diskettes and can be installed within minutes. Distribution disks are copy protected and the program can be installed only three times. (But copies of the program can be uninstalled and then later reinstalled.)

To begin a session in MicroTSP, a data workfile must either be created or loaded (if created in a previous session). When using the CREATE command, a user indicates the frequency of the series (e.g., monthly, quarterly) and the beginning and ending dates. Data can be read into MicroTSP from ASCII, Lotus 123, or Data Interchange Format (.DIF) files. In addition, MicroTSP has been set up to interface smoothly with Citicorp's Citibase database. Data can be stored in two ways. First, the entire data workfile can be saved so that it can be loaded during later sessions. Second, individual series can be stored in database (.DB) files that can be extensively labeled. This labeling feature is useful for documenting data transformations.

There are four basic ways to use MicroTSP: (a) pull down menus and select estimation options, (b) issue a one-word command and respond to prompts, (c) issue a complete command string, or (d) execute a batch file of complete command strings. Clearly, the bias in MicroTSP is toward an interactive style of computing. This does not, however, mean that MicroTSP is not useful for academic research.

MicroTSP is my first choice whenever I want to test the properties of a univariate time series or estimate and diagnose a single equation. In MicroTSP, it is very easy to obtain plots of the autocorrelation, partial autocorrelation, and cross correlation functions along with both the Box-Pierce and Ljung-Box Q statistics. In addition, Dickey-Fuller and Augmented Dickey-Fuller tests (with and without drift and/or trend) can be selected from one of the pull-down menus or can be estimated using the LS (least squares) regression command. The output obtained from the menu-driven ADF test includes MacKinnon's (1991) critical values. MicroTSP also contains an impressive series of tests of structural change and parameter stability (e.g., CUSUM test, CUSUM of squares test, recursive residuals, recursive coefficients), parameter restrictions (e.g., Wald test for multiple linear and nonlinear restrictions, Wald test for multiple linear and nonlinear cross-equation restrictions), and residual features (e.g., LM test for ARCH, White test for heteroskedasticity, Jarque-Bera LM test for normality) that can be applied to single equations estimated via the LS, TSLS (two-stage least squares), or NLS (nonlinear least squares) commands.

Because its programming language is simple (similar to BASIC), I also use MicroTSP to calculate some test statistics not available in other programs. For instance, I recently wrote the some MicroTSP program to compute a variance-ratio statistic (Cochrane 1988; Diebold 1989) to test for fractional integration in two time series, x1 and x2 (see MicroTSP code example below). MicroTSP also can be useful for examining the dynamics of basic "vanilla" vector autoregressive systems. However, there is no lag length test available. Furthermore, there is no test for cointegration in a system, no ability to estimate ARCH models, and no possibilities for Bayesian or frequency domain estimation.

The bias of MicroTSP toward interactive rather than batch use is its biggest strength. For instance, when investigating the time series characteristics of a series, the ability to add lagged values to an ADF regression, estimate it, and then immediately view a plot of the autocorrelation function of the residuals really allows an analyst to get a "feel" for the properties of the data. However, the bias toward interactive use also weakens the program in at least two ways. First, the fact that users can select estimators and diagnostic tests from menus means that novices can produce results mechanically. Second, MicroTSP has no matrix programming capabilities.

Because of these limitations, I use the Regression Analysis of Time Series (RATS) package, developed by Estima, to pick up where MicroTSP leaves off. Unfortunately, RATS is not a particularly user-friendly program. Its manual assumes a degree of methodological sophistication beyond that of most students in an introductory regression course.

RATS can be installed on IBM-compatible personal computers and requires 2 mb of storage space and at least 1 mb of extended memory. There are also versions of RATS for the Macintosh and for mainframe and Unix systems. The program is issued in compressed form on either two 3-1/2 inch diskettes or two 5-1/4 inch diskettes. Installation is straightforward and takes about 15 minutes.

RATS can be used in either interactive or batch mode. When starting RATS in interactive mode, the user enters the RATS Editor, which contains a series of pull-down menus that can be accessed via the keyboard or mouse. These menus are similar to those found in the typical spreadsheet program and are not menus of RATS commands. With these menus, a user can open text, graph, or series editing windows. Text windows are used for editing text and executing complete RATS command strings interactively.

Whether using RATS in interactive or batch mode, a user begins by opening a data file. To do so, CALENDAR and ALLOCATE commands are used to describe the frequency and beginning and ending dates for the data. OPEN and DATA instructions are then issued to identify the input file, the format and organization of the data, and the names of the variables. Data can be read into RATS from ASCII, Lotus 123, dBase, Data Interchange Format (.DIF), or MicroTSP (.DB) files. A utility program called RATSDATA (provided with RATS) can be used to import and export data. A utility program called CITIRATS allows extraction of data from Citicorp's Citibase data base.

RATS is a very powerful program that includes three types of estimation routines. First, the program contains built-in instructions for standard estimators such as OLS, probit, logit, 2SLS, and SUR. Moreover, users can write and maximize their own likelihood functions or invoke RATS instructions for nonstandard Bayesian, Kalman filtering, or frequency domain estimators. Second, the RATS manual includes an impressive array of examples of RATS code for specification tests, residual analysis, and hypothesis tests.

Third, RATS contains a set of PROCEDURES, which are similar to Fortran subroutines. Some of the procedures included with the RATS program are for Dickey-Fuller unit root tests, the Geweke-Porter-Hudak fractional differencing estimator, and Hodrick-Prescott filtering. Because there is a sophisticated and active set of RATS users worldwide, new RATS procedures are written at a rapid pace. Estima facilitates the distribution of these new procedures via a quarterly newsletter; an anonymous ftp site at netec.mcc.ac.uk; and an electronic bulletin board that can be reached at (708) 865-8816. A procedure for the KPSS test (Kwiatkowski, Phillips, Schmidt, and Shin 1992) of the stationarity of a time series was recently posted on the bulletin board. And the May 1994 newsletter contained RATS code for estimating complex GARCH models. Then too, users often make RATS program available. For instance, Giannani (1992, 120-127) provides RATS code for estimating structured impulse response functions with asymptotic confidence intervals. Being able to modify someone else's code rather than starting from scratch is often helpful.

My biggest frustration with Estima and the RATS program is over the mistakes that they make. First, new procedures are sometimes incorporated into RATS without a thorough check of whether they work properly. For example, a co-author and I ran into serious bugs in the GPH.SRC (GPH estimator) and FIF.SRC (long-memory filter) procedures. A second example of their neglect of detail revolves around the formulas used to calculate the Akaike and Schwarz Information Criteria. The RATS manual reports that "there are several equivalent formulations for these criteria." Unfortunately, the formulas they use are NOT mathematically equivalent to the formulas reported in Judge et al. (1985, 242) or Mills (1990, 138).

The RATS programming language is very powerful and is similar to FORTRAN in the way series and vectors are declared, the way variables can be passed along within a program, and the way loops are written. For instance, the RATS program (in the RATS code example below) will replicate Table 8.10 in Mills (1990) by estimating 16 ARMA(p,q) models. RATS also has some useful graphics features associated with its GRAFEDIT utility program (included with RATS). Graphs and figures created in RATS can be imported into a variety of popular graphics and word processing packages.

In some sense, MicroTSP and RATS are very different programs aimed at different audiences. MicroTSP's interactive immediacy is useful when investigating the properties of the data. In contrast, RATS is useful for more complex estimation and programming problems. Fortunately, the weaknesses of one program generally dovetail with the strengths of the other.

References

Cochrane, John H. 1988. "How Big is the Random Walk in GNP?" *Journal of Political Economy.* 96: 893-920.

Diebold, Francis X. 1989. "Random Walks Versus Fractional Integration: Power Comparisons of Scalar and Joint Tests of the Variance-Time Function" in *Advances in Econometrics and Modeling.* ed. Raldev Raj. New York: Kluwer Academic Publishers.

Giannini, Carlo. 1992. *Topics in Structural VAR Econometrics.* Berlin: Springer-Verlag.

Judge, George G., W.E. Griffiths, R. Carter Hill, Helmut Lutkepohl, Tsoung-Chao Lee. 1985. *The Theory and Practice of Econometrics.* 2d ed. New York: John Wiley and Sons.

Kwiatkowski, Denis, P.C.B. Phillips, Peter Schmidt, and Yongcheol Shin. 1992. "Testing the Null of Stationarity Against the Alternative of a Unit Root" *Journal of Econometrics.* 54:159-78.

MacKinnon, James G. 1991. "Critical Values for Cointegration Tests" in *Long Run Relationships: Readings in Cointegration.* ed. Robert Engle and C.W.J. Granger. New York: Oxford University Press.

Mills, Terence C. 1990. *Time Series Techniques for Economists.* New York: Cambridge University Press.

MicroTSP is available (for $295) from Quantitative Micro Software, 4521 Campus Drive, Suite 336, Irvine, CA 92715. Phone: 714-856-3368; FAX: 714-856-2044

RATS, is available from Estima, 1800 Sherman Ave., Suite 612, Evanston, IL 60201. Phone: 800-822-8038; FAX: 708-864-1910. An MSDOS version is available for $420. A version for MS WINDOWS is in final beta test, with availability slated for late 1994 at an approximate price of $500. A UNIX version is also available.

MicroTSP Code Example

```
load jq                                         @loads previously created workfile jq
output b:vr.out                                 @opens the output file b:vr.out
pon                                             @prints output to designated file
smpl 1953.2 1992.4                              @set sample range
genr nobs=@obs(x1)                              @generate nobs = number of observations
genr x1d=x1-x1(-1)                              @generate first difference of x1
genr x2d=x2-x2(-1)                              @generate first difference of x2
for %0 x1 x2                                    @loop to be performed for x1 and then x2
  for !1 = 1 to 8                               @loop to be performed eight times
  smpl 1953.1+!1 1992.4                         @resets the sample for missing observations
  genr df!1=(nobs-!1+1)                         @begin variance ratio calculation
  genr diff!1=(%0-%0(-!1)-((!1)*(@mean(%0d))))
  genr sqdiff!1=diff!1*diff!1
  genr sums!1=@sum(sqdiff!1)
  genr sigma!1=(1/df!1)*(sums!1)
  genr vratio!1=((!1)*(sigma1))/(sigma!1)       @finish variance ratio calculation
  cova vratio!1                                 @yields descriptive statistics for vratio!1
  next                                          @returns to "for !1" statement
next                                            @returns to "for %0" statement
exit                                            @exit to DOS
```

RATS Code Example

```
calendar 1950                                         *defines begin date for annual data
allocate 85:1                                         *defines end date
open data b:mills.dat                                 *opens file with ASCII data
data(format=free,org=obs) / obsno income consump      *reads in data
set newinc 50:1 85:1 = income/100                     *transformed income = newinc
table                                                 *yields statistics for all series
smpl 53:1 85:1                                        *resets sample period
dofor p = 0 to 3                                      *begin loop for AR parameters
  dofor q = 0 to 3                                    *begin loop for MA parameters
    boxjenk(constant,ar=p,ma=q,iters=50) newinc       *estimate ARMA(p,q) model for newinc
    display @20 'AR(p)' @35 'MA(q)'                   *display labels
    display @20 ##.## p @35 ##.## q                   *display values of p and q
    compute akaike = $                                *compute AIC, $ means line continues
      log(%rss/(%nobs-(p+q-1)))+((p+q)*2.0)/%nobs
    compute schwarz = $                               *compute SIC, $ means line continues
      log(%rss/(%nobs-(p+q-1)))+((p+q)*log(%nobs))/%nobs
    display @20 'AKAIKE' @35 'SCHWARZ'                *display labels
    display @20 ###.### akaike @35 ###.### schwarz    *display AIC and SIC
  end dofor                                           *returns to "dofor q" statement
end dofor                                             *returns to "dofor p" statement
halt                                                  *finish, exit to DOS
```

Political Analysis Ad

Background: Statement on Statistical Reporting, Archiving and Replication

Walter R. Mebane, Jr.
Cornell University

The "Statement on Statistical Reporting, Archiving and Replication: Norms for Publication" grew out of discussions that have taken place over the past few years at several gatherings of political methodologists. A visit by then-new ICPSR director Richard Rockwell at the 1992 Political Methodology Section business meeting (held during the APSA Annual Meetings) prompted an informal suggestion from the floor that authors be encouraged to deposit at ICPSR any data used to produce published quantitative analysis. ICPSR had recently created a Class 5 category to support deposits of data from economists working with National Science Foundation support, who were as a rule expected to submit data to a public archive to fulfill the contract with NSF. Following this meeting, *Political Analysis* editor John Freeman adopted a version of the data-deposit policy, effective beginning with volume 4 of the journal. At the 1993 summer conference on political methodology held at Florida State University in Tallahassee there was a suggestion that a committee be formed to discuss standards for how results from statistical analysis ought to be reported in published work. There was widespread agreement that current conventions for reporting left much to be desired—especially, crucial details needed to evaluate the results are too often omitted—but there were significant disagreements about what (if any) standards ought to be enforced. Opinions varied, for instance, on the merits of summary statistics such as R^2 in regression-type analysis.

These and other discussions prompted Political Methodology Section President Larry Bartels during the Fall of 1993 to form an ad hoc committee to study the reporting problem. The committee would report and solicit comments first at the 1994 summer conference (at the University of Wisconsin in Madison) and then on a panel at the 1994 APSA Annual Meeting. The committee's report would then be formally presented to and voted on by the Section at the 1994 business meeting. Discussion at the summer meeting included comments by *American Journal of Political Science* editor Ken Meier regarding his good experience with a policy of requiring that authors whose work is to be published in the *Journal* include a footnote stating that the data used in the article are available from the author and how the data may be obtained. At APSA a lively discussion led to a move at the business meeting to amend one point in the committee's original statement. Instead of the current provision that authors must make data and documentation available through a public archive (such as ICPSR), the original statement had proposed to require authors to deposit a copy of the materials with the journal editor along with the final version of the accepted manuscipt; responsibility for distributing data on request was to be the author's (possibly using a public archive); the editor would do no more than file the submitted materials along with the other materials that are routinely preserved for each published piece. The move to amend the original statement passed by a substantial majority. The amended statement received unanimous support.

Editors of several of the most prominent journals in political science are adopting versions of the data-availability policy outlined in the "Replication" section of the statement. Most are using Ken Meier's policy for *AJPS*. Comments and suggestions regarding the statement or related issues may be directed to Larry Bartels or to any of the ad hoc committee members, but the most fruitful approach may be to get in touch with the editor or editorial board of the journal of greatest interest to you. The near future looks bright for these improvements in practices for reporting the results of quantitative research in political science.

Statement on Statistical Reporting, Archiving and Replication: Norms for Publication

James Stimson (chair)
University of Minnesota

Charles Franklin
University of Wisconsin

Walter R. Mebane, Jr.
Cornell University

Philip A. Schrodt
University of Kansas

B. Dan Wood
Texas A&M University

We propose as a principle that the materials that are crucial to statistical inference must to the greatest degree possible be available to readers for review.[1]

[1] **Editor's Note:** The authors of this statement constitute the Committee on Reporting and Replication, Political Methodology Section of the American Political Science Association. This statement was delivered and discussed at the 1994 Annual Meeting of the American Political Science Association, New York, September 1–4.

Reporting:

An important application of this principle is that statistical estimates such as regression coefficients should always be reported with some measure of dispersion. Of those in general use—standard errors, t statistics, probability values, and asterisks for tests of significance—the latter two are clearly least desirable because information is lost through standardization and assumptions about the test. Standard errors and t statistics may each be computed from knowledge of the other (and the coefficients). Preferences between the two are thus more a matter of convention than substance. It is desirable to have a convention to avoid misinterpretation and confusion. We propose the standard error as such a convention.

Replication:

Perhaps the single most serious deficit in the current practice of political science is that published work is rarely replicated. The result is that inferences are rarely challenged. In the small number of cases where they are, design and analysis are typically so different between studies as to rule out knowing whether differing conclusions can be compared. Even with widespread use of archival data, replication is typically so tedious that it is wasteful of scientific talent and quite often it is not possible. Where data are held privately, public claims are not subject to disconfirmation at all.

We identify two deficiencies in normal practice: (1) we do not make available records of research decisions, indeed typically we do not keep adequate records of them for our own use; and (2) we do not share the data on which inferences rest.

To remedy these deficiencies we propose that manuscripts accepted for journal publication in political science be accompanied by some compact form of research documentation, including, where appropriate, original data. Authors would be required to retain an identical copy of the submitted materials. Such documentation and data would be limited to materials used in the article. Each article would carry a note informing readers how they may obtain the data and documentation necessary to reproduce the reported findings exactly. The materials must be conveyed through a public archive. Publicly available data would not be reported; documentation for data management and analysis would be sufficient in this case.

Exceptions would be allowed in some circumstances at the discretion of the editor. We identify three specific cases: (1) where public access to data may harm research subjects; (2) where public access may violate agreements between researcher and subject as to confidentiality (or preclude the possibility of research that requires confidentiality); (3) where the author has recently produced an original data collection and has had little opportunity to explore it. The journal editor would be the final arbiter of the issue.

In the case of original data collections we propose that exceptions be rare rather than common. This recommendation is based on the limited scope of the reporting requirement: only data used to compute statistics reported in the article must be documented. We also have in mind the typically lengthy span between data collection and final publication. We propose a five-year limitation on claims of private data ownership. The five-year period would be counted starting at the time when collection of the data used in the article was completed. In such cases where private data are withheld for further use, the author's note would provide a specific date on which the data would become available.

Notes

The motivation for requiring data and documentation to be deposited along with the submitted manuscript is that it requires documentation to be created at a moment when the material is fresh, not months or years later when the author is no longer capable of adequate documentation. The author who submits materials, e.g. a diskette, may easily make an identical copy to set aside for future requests.

We anticipate that instructional use of such data may have the beneficial impact of making routine the notion that important research will be replicated. This should cause each of us to focus on replicability at the time the work is done, and accordingly improve the routines by which we work.

The proposal is deliberately vague about the form of deposited data and documentation. We recognize differences in procedures and equipment as well as the likelihood of continuing technological change.

The proposed requirements apply only to work that makes empirical claims with numeric data or estimates. The criterion in practice comes down to whether the work is exactly replicable in principle or not. This is readily signaled by numeric tables and figures or by reports of estimated parameters.

We seek to encourage greater publication of data for small sample research where space constraints permit. Data tables that support the published estimates would be of great value to the scholarly community.

Replication Datasets To be Listed in *TPM*

Gary King
Harvard University

An increasing number of political science journals, book presses, and granting organizations are requiring authors to add a footnote to publications indicating in which public archive they will deposit the information necessary to replicate their numerical results, and the date when it will be submitted (or an explanation if the data could not be archived). In order to encourage these contributions to the scholarly community, *The Political Methodologist* will provide authors some additional visibility by listing a brief citation to their "replication dataset," and the corresponding publication for which it was created, in our next available issue.[1]

Two of the archives willing to accept replication datasets include the *Social Science Research Archive* collection of the *Public Affairs Video Archive* (PAVA) at Purdue University and the "Class V collection" at the *Inter-University Consortium for Political and Social Research* (ICPSR) at the University of Michigan. Both archives will forever keep and distribute replication datasets and make them known to others. In order to submit the data, put it on a disk or tape and mail it to PAVA (Director; Public Affairs Video Archive; Purdue University; 1000 Liberal Arts Building, West Lafayette, Indiana 47907-1000) and/or the ICPSR (Director, User Support; ICPSR; P.O. Box 1248; Ann Arbor, MI 48106). An easier approach is to put your data in a self-extracting archive file (with a utility such as PKZIP for the DOS operating system, TAR for Unix, or StuffIt for the MacIntosh) and submit it via anonymous FTP; you should also announce the file name, and article, book, or dissertation citation in an accompanying electronic mail message. To send to PAVA, FTP to `pava.purdue.edu` in directory `pub/incoming` and send electronic mail to `info@pava.purdue.edu`. To submit to the ICPSR, FTP to `ftp.icpsr.umich.edu` in directory `pub/incoming` and send electronic mail to `jan@tdis.icpsr.umich.edu`.

Domestic Policy Mood: An Update

James Stimson
University of Minnesota

This is the second release of the domestic policy mood time series. It follows publication four years earlier in Stimson (1991). This note documents revisions in data and estimation procedures.

Largest differences arise from the availability of an additional four years of preference data along with minor additions of previously fugitive series from the period 1952-1990. Two revisions of estimation method are worthy of note. Prior estimates were the product of a backward recursion algorithm. These estimates are produced with a hybrid that performs both backward and forward recursions on each iteration and then averages the two independent series. The product moment correlation of the two series (on the final iteration) then becomes a reliability estimate. For the series at hand, these estimates are .857 (annually) and .860 (quarterly).

The second change in procedure is the introduction of a smoothness prior and explicit smoothing in the iterative estimation. Since the survey data which are input are known to have sampling error, a smooth approximation to the data can be a more valid measure of the concept. Smoothing is implemented by estimation of α in the exponential smoothing model: $\hat{y}_t = \alpha y_t + (1-\alpha)\hat{y}_{t-1}$. The criterion is minimization of one step forecast error within sample. An important side benefit of smoothing is that the annual series can be extended further backward in time (to 1952) and a lengthy quarterly series becomes possible. In both cases smoothing tames the large variation that would otherwise occur when sparse observations are weighted at 1.0. The estimation method always gains reliability by using more observations for estimating each point in the series, but at the cost of large time intervals. Evidence of the good effect of smoothing is seen now in how closely the series estimated with fewer observations and finer time intervals track, for example, biennial series (not shown).

Input series are scored as percent liberal / (percent liberal + percent conservative). Neutral and uncodable responses are not used. Metric information is lost during estimation and then reintroduced after. After communalities are known, the final scale is a weighted average (by communality) of the means and variances of the items of which it is composed. It is thus interpretable in an approximate sense as the average percent of respondents giving liberal responses in any given year.

Presented here, in Tables 1 and 2, are the estimates through 1993:4. The time spans chosen are those for which I am confident that the data and technology are sufficient. In the quarterly case the starting date, 1961:1, is chosen to maximize the length of the series while minimizing the number of quarters in which no observations are available.

Electronic (ASCII) copies of these series and expected future updates are available by email from <stimson@atlas.socsci.umn.edu>.Refere

Stimson, James A. 1991. *Public Opinion in America: Moods, Cycles, and Swings.* Boulder, CO: Westview Press.

[1] **Editor's Note:** This will be a standing policy of *The Political Methodologist* from this issue forward.

Table 1: Domestic Policy Mood:
Annual, 1952 to 1993

Period	Mood	Period	Mood
1952	55.89	1973	63.14
1953	59.48	1974	63.31
1954	54.44	1975	61.98
1955	62.30	1976	58.59
1956	63.57	1977	57.40
1957	64.70	1978	56.33
1958	66.56	1979	55.29
1959	67.95	1980	54.25
1960	67.78	1981	54.90
1961	71.76	1982	56.55
1962	71.34	1983	59.36
1963	70.23	1984	60.79
1964	67.27	1985	60.92
1965	64.52	1986	62.45
1966	64.92	1987	62.91
1967	65.72	1988	65.54
1968	64.15	1989	67.59
1969	61.39	1990	68.10
1970	63.61	1991	68.34
1971	64.83	1992	67.15
1972	64.31	1993	66.61

Table 2: Domestic Policy Mood:
Quarterly, 1961:1 to 1993:4

Period	Mood	Period	Mood	Period	Mood	Period	Mood
1961:1	74.67	1969:2	61.83	1977:3	60.08	1985:4	61.02
1961:2	72.91	1969:3	64.61	1977:4	59.58	1986:1	61.61
1961:3	73.59	1969:4	65.52	1978:1	58.92	1986:2	62.37
1961:4	73.51	1970:1	66.33	1978:2	57.83	1986:3	60.87
1962:1	73.20	1970:2	66.85	1978:3	57.14	1986:4	62.91
1962:2	72.26	1970:3	67.48	1978:4	56.75	1987:1	64.11
1962:3	71.61	1970:4	66.31	1979:1	56.83	1987:2	64.62
1962:4	71.58	1971:1	65.53	1979:2	57.13	1987:3	64.44
1963:1	71.65	1971:2	68.72	1979:3	57.36	1987:4	64.25
1963:2	71.91	1971:3	69.34	1979:4	56.26	1988:1	64.69
1963:3	70.93	1971:4	69.21	1980:1	55.27	1988:2	64.39
1963:4	69.61	1972:1	68.76	1980:2	55.25	1988:3	63.46
1964:1	66.59	1972:2	67.85	1980:3	55.27	1988:4	63.56
1964:2	65.60	1972:3	68.54	1980:4	55.31	1989:1	65.37
1964:3	64.69	1972:4	68.18	1981:1	54.38	1989:2	64.82
1964:4	65.34	1973:1	66.96	1981:2	53.94	1989:3	66.16
1965:1	62.57	1973:2	67.01	1981:3	54.96	1989:4	66.58
1965:2	64.95	1973:3	67.33	1981:4	55.97	1990:1	66.44
1965:3	63.61	1973:4	65.95	1982:1	56.46	1990:2	65.92
1965:4	63.04	1974:1	65.96	1982:2	57.43	1990:3	64.62
1966:1	63.63	1974:2	64.86	1982:3	58.86	1990:4	65.55
1966:2	64.44	1974:3	65.43	1982:4	59.60	1991:1	66.23
1966:3	60.54	1974:4	64.66	1983:1	60.09	1991:2	65.81
1966:4	61.82	1975:1	63.97	1983:2	59.52	1991:3	64.21
1967:1	62.35	1975:2	63.16	1983:3	59.57	1991:4	64.69
1967:2	60.45	1975:3	62.53	1983:4	60.28	1992:1	65.11
1967:3	61.66	1975:4	61.17	1984:1	61.23	1992:2	65.37
1967:4	62.04	1976:1	60.94	1984:2	60.89	1992:3	65.77
1968:1	63.04	1976:2	60.60	1984:3	63.56	1992:4	64.32
1968:2	63.54	1976:3	60.24	1984:4	63.67	1993:1	64.39
1968:3	63.71	1976:4	59.90	1985:1	62.18	1993:2	64.41
1968:4	63.95	1977:1	60.01	1985:2	62.20	1993:3	63.10
1969:1	62.63	1977:2	59.98	1985:3	60.71	1993:4	62.57

Exogeneity, Inference, and Granger Causality: Part II: The Case of Integrated Regressors

Jim Granato
Michigan State University

Renée M. Smith
University of Rochester

This is the second of a two-part discussion of Granger causality tests and issues of exogeneity and inference in time series analysis. In Part II, we assume that the variables under consideration are integrated processes that can be made stationary by taking Rrstdifferences.

One of the key questions addressed in our Part I discussion was: Under what conditions can the processes that generated the independent variables in a single-equation model be ignored without losing information relevant to the consistent and efficient estimation of the parameters in the conditional model? The answer when regressors are stationary is that the conditional model can be estimated separately provided that weak exogeneity holds. In Part II, we address this same issue within the context of integrated regressors. Although weak and strong (or strict) exogeneity continue to be important concepts, when time series are integrated, these concepts have implications that extend beyond the issue of consistency and instead bear on the distribution of certain coefficients in the conditional model.

To begin to show why this is true, consider the following time series process:

$$x_t = \delta + \rho x_{t-1} + v_t \tag{1}$$

where $\rho = 1$ and where v_t has a mean of zero and a variance of σ_v^2. Because $\rho = 1$, the effects of random shocks to x_t persist rather than die out as t approaches infinity.[1] This time series also can be expressed in terms of partial sums. For instance, starting with an initial condition of x_0 and summing over the process in Equation [1] yields:

$$x_t = x_0 + \sum_{i=1}^{t} v_i \tag{2}$$

Notice that $\sum_{i=1}^{t} v_i$ represents the sum of the stochastic deviations from $\rho = 1$.[2] Assuming that $x_0 = 0$ (without loss of generality), the variance of this integrated process is time-dependent and is given by $t\sigma^2$. The intuition to retain from these definitions is that when a time series is a first-order integrated process (denoted I(1)), the stochastic errors accumulate over time and do not disappear as the sample size approaches infinity. Furthermore, *any* information (any AR or MA components) contained in v_t will accumulate over time.

The fact that integrated processes are a cumulative function of past information has important implications for the distribution theory that applies to coefficient estimates and test statistics obtained from models that contain such regressors. These consequences can be illustrated by discussing the large-sample convergence and distribution results that apply in the following context:

$$y_t = \beta x_t + \epsilon_t \tag{3}$$

$$x_t = x_{t-1} + v_t \tag{4}$$

where v_t has a mean of zero but is not necessarily iid, $x_t \sim I(1), y_t \sim I(1)$, and ε_t is a stationary, mean zero process.[3] Furthermore, Ω, the "long-run" variance-covariance matrix of the disturbance vector, $u_t\prime = (\varepsilon_t, v_t)$, is given by:

T

$$\Omega = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \tag{5}$$

In this case, Equation [3] represents a structural, conditional model and Equation [4] is a marginal model. Engle and Granger (1987) call Equation [3] a cointegrating regression because the stochastic trend (associated with the unit roots in each univariate series) is common to both y_t and x_t. However, these individual unit roots cancel each other so that the linear combination, $\hat{\varepsilon}_t = y_t - \hat{\beta} x_t$, is stationary (denoted I(0)). Asymptotic Theory

Although the notation may appear complex or unusual, it is quite instructive to examine the asymptotic properties and limiting distribution for $\hat{\beta}$, the OLS estimator of β in Equation [3]. Park and Phillips (1988) derive the following results:

$$\frac{1}{T}\sum_{t=1}^{T} x_t \epsilon_t = \frac{1}{T}\sum_{t=1}^{T} S_{2t}\epsilon_t \Longrightarrow^d \int_0^1 B_2 dB_1 + \sigma_{21} \tag{6}$$

[1] When $\delta = 0$ and v_t is identically and independently distributed (iid), then x_t is often called a random walk. When $\delta = 0$ and v_t is serially correlated and weakly dependent (e.g.; contains either AR or MA components), then this process is simply said to be integrated. In either case, if $\delta \neq 0$, the process is said to have drift. In this paper, we will assume each series has no drift.

[2] Thus, x_t is said to contain a stochastic trend that arises because x_t has a unit root in its autoregressive representation. A time series also can contain a deterministic trend or can have both a deterministic and a stochastic trend. Park and Phillips (1988, 1989) provide the asymptotic theory for such regressors. We will not discuss times series with deterministic trends and drift in this paper.

[3] If ε_t is not a mean zero process, it can be made so by including a constant term in the regression. Throughout this text, the constant term is dropped for notational convenience.

$$\frac{1}{T^2}\sum_{t=1}^{T} x_t^2 = \frac{1}{T^2}\sum_{t=1}^{T} S_{2t}S_{2t} \Longrightarrow^d \int_0^1 B_2 B_2 \qquad (7)$$

where $\Longrightarrow^d$ indicates weak convergence in distribution and where B_1 and B_2 are the Brownian motion processes associated with $S_{1t} = \sum_{i=1}^{t} \varepsilon_i$ and $S_{2t} = \sum_{i=1}^{t} v_i$, the partial sum processes for ε_t and v_t, respectively.[4]

There are two key pieces of intuition to draw from these results. First, notice that the regressor-disturbance correlation in [6] does not converge to zero but is confounded by two types of bias: (a) the dependence between the stochastic trends in y_t and x_t and (b) the correlation between x_t and ε_t that arises when v_t is not iid but contains AR or MA components.[5] Second, note that when regressors are integrated, this nonzero covariance is of no asymptotic consequence; $\hat{\beta}$ is estimated consistently regardless. This is true because as shown in expressions [6] and [7], the magnitude of the sequence of the sum of squares is $O(T^2)$ and is larger than the $O(T)$ magnitude of the regressor-disturbance correlation. Since these two sequences converge at different rates, the signal from the variance of x_t swamps out the sample correlation between x_t and ε_t.[6]

Despite the large-sample consistency of $\hat{\beta}$ in the face of many types of simultaneous equation or omitted variable biases, the variance of this OLS estimator is biased in such a way that $\hat{\beta}$ has a non-centered, nonstandard limiting distribution given by:

$$T(\hat{\beta} - \beta) \Longrightarrow^d \left(\int_0^1 B_2 B_2\right)^{-1}\left(\int_0^1 B_2 dB_1 + \sigma_{21}\right) \qquad (8)$$

Because the distribution of $\hat{\beta}$ is confounded by unit-root dependencies and by the nuisance parameter σ_{21}, the usual t-ratios and F statistics do not apply, and hence, drawing inferences about $\hat{\beta}$ becomes difficult.[7] Note, however, that if the confounding effects due to serial correlation in v_t are controlled for, the only remaining biases in the limiting distribution of $\hat{\beta}$ are from the joint endogeneity of y_t and x_t.[8] When the dependence between y_t and x_t due to cointegration is ignored during estimation, the unit root components of these series are estimated and, as a result, $\hat{\beta}$ has a "unit-root" distribution that must be computed in different multivariate contexts.

However, as Phillips (1987) proves, when both the correlation between y_t and x_t due to cointegration and the endogeneity of x_t due to serial correlation are estimated or controlled, the limiting distribution of $\hat{\beta}$ is:

$$T(\hat{\beta} - \beta) \Longrightarrow^d \left(\int_0^1 B_2 B_2\right)^{-1}\left(\int_0^1 B_2 dB_{1\mathbb{E}}\right) \qquad (9)$$

where $B_{1\mathbb{E}}$ is the Brownian motion process associated with the partial sum of the conditional variance of ε_t given v_t. In other words, $B_{1\mathbb{E}}$ is the Brownian motion process B_1 purged of its dependence on the Brownian motion process B_2. Once both types of biases are removed, the distribution of $\hat{\beta}$ becomes a mixture of *independent* normally distributed components, and hence, $\hat{\beta}$ is asymptotically normally distributed (Phillips and Loretan 1991).

Exogeneity, Inference, and Granger Causality

Understanding the intuition behind these asymptotic results is immediately relevant to understanding issues of inference and exogeneity as well as the use of Granger causality tests within the context of integrated and cointegrated regressors. For instance, in light of the above discussion, we can determine under what conditions the marginal model for x_t can be ignored without losing information relevant to inference. The answer in the case of integrated regressors is that the conditional model can be separately estimated with inference proceeding on the basis of the usual t-ratios and F statistics provided that strong (i.e.; strict) exogeneity holds. Strict exogeneity holds when $\sigma_{21} = \sigma_{12} = 0$, and this is precisely the condition under which the likelihood functions for the marginal and conditional models are independent and under which $\hat{\beta}$ in Equation [4] is normally distributed.

Because strict exogeneity is an important condition underlying single-equation estimation with integrated and cointegrated time series, analysts may want to use Granger

[4] Intuitively, Brownian motion is a continuous-time version of a discrete-time random walk. More formally, a Brownian motion process, $B(t)$, is continuous but nowhere differentiable; has stationary, independent increments that are normally distributed random variables; and therefore, is distributed normally with a mean of zero and an unbounded variance that is time dependent. By the functional central limit theorem, when v_i is either iid$(0, \sigma_v^2)$ or alpha-mixing with variance equal to σ_v^2 (e.g.; serially correlated disturbances are α-mixing), then $X_n(t) = N^{\frac{1}{2}} \sum_{i=1}^{[nt]} v_i$ converges weakly in distribution to Brownian motion (Phillips and Durlauf 1986). (The integer that is closest to n is denoted $[nt]$.) While the usual central limit theorem results (e.g.; Lindberg-Levy, etc.) tell us that the endpoint of a sequence is normally distributed, the *functional* central limit theorem implies that this normality holds for all subsets of the sample as of today.

[5] The former dependence generates the terms $\int_0^1 B_2 dB_1 + \sigma_{21}$ where the nuisance parameter σ_{21} arises because $E(v_{t-p}\varepsilon_t) \neq 0$ for all p leads and lags, (i.e.; for $p \neq 0$). The latter correlation contributes to the term σ_{21} because it leads to a situation where $E(v_{t-p}\varepsilon_t) \neq 0$ for all $p \geq 0$.

[6] As shown in the expression in [8] below, another implication of these different rates of convergence is that $\hat{\beta}$ converges to β more rapidly (at rate T) than in the stationary case (Stock 1987).

[7] These dependencies can potentially arise whether $\hat{\beta}$, the long-run effect of x_t on y_t, is estimated using a single cointegrating regression or a single-equation error correction model (Phillips 1988, 1991; Urbain 1992).

[8] In fact, the logic of estimating and removing the effects of the endogeneity of the regressors caused by serial correlation to obtain an estimator with a "unit-root" distribution undergirds many of the proposed tests for unit roots in a univariate series (e.g.; Dickey and Fuller 1981; Phillips and Perron 1988).

causality tests to test their exogeneity assumptions. The primary complication that arises for Granger causality tests when regressors are integrated revolves around the appropriate asymptotic distribution of the test statistics. Recall that most analysts implement Granger causality tests within the context of a vector autoregression model in which the level of each variable in a system is regressed on the lagged levels of all of the other variables in the system. However, using a VAR in levels to perform Granger causality tests is problematic because when this estimator is used, the unit roots in each time series are estimated. Therefore, if the regressors are integrated but not cointegrated, Wald tests on blocs of coefficients will have nonstandard limiting distributions (Sims, Stock, and Watson 1990). In fact, even when integrated regressors are cointegrated, Wald tests of blocs of coefficients will often have nonstandard limiting distributions because information about the number of unit roots in the system is not being used.[9]

One way to avoid nonstandard "unit root asymptotics" is to implement Granger causality tests within the context of a system of error correction equations (Engle and Granger call this a restricted, diagnostic VAR). For instance, consider the following system in which there is a single cointegrating relationship:

$$\begin{aligned} \Delta x_t &= \Theta_{10} + \sum_{p=1}^{P} \delta_{1p} \Delta y_{t-p} \\ &\quad + \sum_{p=1}^{P} \gamma_{1p} \Delta x_{t-p} \\ &\quad - \alpha_1 (x_{t-1} - \beta_1 y_{t-1}) + u_{1t} \end{aligned} \tag{10}$$

$$\begin{aligned} \Delta y_t &= \Theta_{20} + \sum_{p=1}^{P} \delta_{2p} \Delta y_{t-p} \\ &\quad + \sum_{p=1}^{P} \gamma_{2p} \Delta x_{t-p} \\ &\quad - \alpha_2 (y_{t-1} - \beta_2 x_{t-1}) + u_{2t} \end{aligned}$$

where u_{1t} and u_{2t} are each iid and mean zero and where the model for Δy_t represents a complete dynamic specification. Within this context, the null hypothesis that y_t does not Granger cause x_t can be expressed as $H_0 : \alpha_1 \beta_1 = 0$ and $\sum_{p=1}^{P} \delta_{1p} = 0$ (Engle and Granger 1987; Toda and Phillips 1991). Toda and Phillips (1991) discuss the nuances of the distribution theory for Granger causality tests within an ECM framework and show that it is generally easier to avoid "unit-root" asymptotics in a multivariate ECM context because in an ECM, the unit roots in y_t and x_t are eliminated by construction.

Provided that y_t and x_t are cointegrated, a finding of Granger noncausality from y_t to x_t (i.e.; a failure to reject the null) is consistent with an assumption that x_t is strictly exogenous to y_t. When strict exogeneity fails, multivariate system estimators (Johansen 1988; Stock and Watson 1988) must be used to obtain normally distributed estimates of $\hat{\beta}$ (Phillips 1991). In fact, Phillips (1988) shows that FIML system estimation is generally necessary if v_t contains moving average components. In addition, system estimation becomes important when more than one cointegrating relationship exists.

Conclusion

The purpose of this paper has been to highlight the problems of inference that arise when regressors are integrated and cointegrated. Although focusing on these problems may seem pessimistic, we think that the literature discussing the estimators derived to solve these problems is more accessible once the intuition behind the above results is understood.

References

Dickey, D.F. and W.A. Fuller. 1981. "Likelihood Ratio Statistics for Autoregressive Time Series With a Unit Root" *Econometrica*. 49: 1057–1072.

Engle, R.F. and C.W.J. Granger. 1987. "Co-integration and Error Correction: Representation, Estimation and Testing" *Econometrica*. 55: 251–276.

Johansen, S. 1988. "Statistical Analysis of Cointegrating Vectors" *Journal of Economic Dynamics and Control*. 12: 231–254.

Park, J.Y. and P.C.B. Phillips. 1988. "Statistical Inference in Regressions with Integrated Processes: Part I" *Econometric Theory*. 4: 468–497.

Park, J.Y. and P.C.B. Phillips. 1989. "Statistical Inference in Regressions with Integrated Processes: Part II" *Econometric Theory*. 5: 95–131.

Phillips, P.C.B. 1987. "Time Series Regression with Unit Roots" *Econometrica*. 55: 277–302.

Phillips, P.C.B. 1988. "Reflections on Econometric Methodology" *Economic Record*. 64: 344–359.

Phillips, P.C.B. 1991. "Optimal Inference in Cointegrated Systems" *Econometrica*. 59: 283–306.

[9] Sims, Stock, and Watson (1990) seem to indicate that, when integrated regressors are cointegrated, the Wald test of blocs of coefficients on lagged levels regressors have their usual asymptotic χ^2 distribution when the long-run relationship involves the variable that is excluded under the null hypothesis. However, Toda and Phillips (1991) show that this result does not hold when more than one variable is excluded under the null hypothesis unless the excluded variables are themselves sufficiently cointegrated so that tests of blocs of zero restrictions do not involve integrated but noncointegrated regressors.

Phillips, P.C.B. and S.N. Durlauf. 1986. "Multiple Time Series Regression with Integrated Processes" *Review of Economic Studies.* 53: 473–495.

Phillips, P.C.B. and M. Loretan. 1991. "Estimating Long-run Economic Equilibria" *Review of Economic Studies.* 58: 407–436.

Phillips, P.C.B. and P. Perron. 1988. "Testing for a Unit Root in Time Series Regression" *Biometrika.* 75: 335–346.

Sims, C., J. Stock, and M. Watson. 1990. "Inference in Linear Time Series Models with Some Unit Roots" *Econometrica.* 58: 113–144.

Stock, J. 1987. "Asymptotic Properties of Least-Squares Estimators of Cointegrating Vectors" *Econometrica.* 55: 1035–1056.

Stock, J. and M. Watson. 1988. "Testing for Common Trends" *Journal of the American Statistical Association.* 83: 1097–1107.

Toda, H.Y. and P.C.B. Phillips. 1991. "Vector Autoregressions and Causality" Cowles Foundation Discussion Paper No. 977.

Urbain, J.P. 1992. "On Weak Exogeneity in Error Correction Models" *Oxford Bulletin of Economics and Statistics.* 54: 187–207.

GAUSS Code For Re-sampling Cases and Residuals in Bootstrapping a Regression Model

Christopher Z. Mooney
West Virginia University and University of Essex

In bootstrapping a statistical model, it is vital that the analyst re-sample from the random component of that model (Freedman 1981). In a regression model, this random component can be embedded in either the residuals or the case-wise structure of the data. This depends on whether the data were generated using fixed independent variables, as in a classic experimental design, or by a design where the independent and dependent variables were equally subject to random influences, as in a sample survey. In the former case, it is proper to re-sample the residuals, while in the latter it is proper to re-sample the data case-wise (Mooney and Duval 1993, 15-20).

The following GAUSS code will allow for the re-sampling of either cases or residuals, depending on how the flag, "resamp," is set in the first line. The regression proc, "regress," was written locally, but the standard proc, "ols," can be used. Running this code will result in a (Bx1) matrix, "bb2," of bootstrapped regression coefficients. This vector can then be manipulated in the ways described in Mooney and Duval (1993) to make inferences about the slope.[1]

For an ASCII version of a complete program to undertake this procedure, e-mail me at CMOONEY@ESSEX.AC.UK.

References

Freedman, D.A. 1981. "Bootstrapping Regression Models." *Annals of Statistics.* 9:1218-1228.

Mooney, C.Z, and Duval, R.D. 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-095. Newbury Park, CA: Sage.

```
GAU::

resamp = 2; @ Flag to set re-sampling
    type: 1 = residuals,
    2 = cases@
i=1; B = 1000; @i = the looping index;
    B = the number of
    bootstrap re-samples @
bb2 = zeros(B,1); @Setting up empty vector
    to hold bootstrapped
    beta-hat's@


@GIVEN A VECTOR OF OBSERVED X AND Y, AND
A VECTOR OF RESIDUALS, "resid," COMPUTED BY
REGRESSING Y ON X, THEN:@

    @ RESAMPLING@

index=seqa(1,1,N); @Setting up index to be
   re-sampled@
do while i<=B; @Starting the re-sampling
   loop@
   @Setting up cases re-sampling
   index@
cindex=submat(index,ceil(rndu(N,1)*N)',0);
xc=x[cindex,.];
yc=y[cindex,1]; @Re-sampling cases @

   @Re-sampling residuals@
re=submat(resid,ceil(rndu(N,1)*N)',0);
yr=x*bet+re; @Setting up re-sampled Y
   vector for
   residuals re-sampling@

if resamp==1; @Responding to re-sampling
   flag@
```

[1] Thanks to Bob Duval and Burak Saltoglu for their help developing these algorithms.

```
{beta,se,r2}= regress(x,yr); @Residuals
   re-sampling
   regression@

else; @Responding to re-sampling flag@

{beta,se,r2}=regress(xc,yc); @Cases re-sampling
   regression@

endif; @End of loop choosing which type
   of re-sampling@

bsb2=beta[2,.]; @Identifying the regression
   coefficient@

bb2[i,1] = bsb2; @Setting up vector of
   bootstrapped
   beta-hats@

i=i+1; @Adds 1 to bootstrap looping index@

endo; @End of resampling loop@
```

Constructing Bootstrap Confidence Intervals Using GAUSS

Christopher Z. Mooney
West Virginia University and University of Essex

The construction of confidence intervals from a vector of bootstrapped statistics can proceed in one of several ways, depending upon the amount of structure the analyst wishes to impose on the data, the suspected level of skew of the sampling distribution of the statistic, the level of complexity of the estimator, or other reasons (Mooney and Duval 1993, 33-42; Efron and Tibshirani 1993, 153-199). Perhaps the most conservative procedure is to develop several types of these intervals, and check them for congruence.

The GAUSS code below will develop the normal approximation, percentile, BCa, and percentile-t bootstrap confidence intervals around a mean.[1]The code can be modified in a straightforward fashion to develop intervals around any parameter of interest.

For an ASCII version of a complete program to undertake this procedure, e-mail me at CMOONEY@ESSEX.AC.UK.Refere

Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Mooney, C.Z, and Duval, R.D. 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-095. Newbury Park, CA: Sage.

[1]Thanks to Bob Duval and Burak Saltoglu for their help developing these algorithms.

GAU::

```
@Setting up the preliminaries@
tcal=1.9C; @t-score for normal
   approximation CI@
B=1000; @Number of bootstrap
   re-samples@
i=1 ; j=1; @Set indices for re-sample, and
   jackknife for BCa@
prop=0; @Set proportion base for z0
   of BCa@
x_bar_bt = zeros(B,1); @Define empty
   vector for x-
   bar*'s@
sd_xb_bt = zeros(B,1); @Define empty
   vector for SD's
   of x-bar*'s@
t_boot = zeros(B,1); @Define empty vector
   for standardized x-bar*'s@
mn_jack = zeros(n,1); @Define empty vector
   for jackknifed means@

@GIVEN A VARIABLE, x, OF SAMPLE SIZE, n,
CONSTRUCTING 95% CONFIDENCE
INTERVALS AROUND ITS MEAN:@

@Calculating the sample mean of x for the
full sample@
x_bar = meanc(x);
sd_x = stdc(x);
sd_x_bar = sd_x/sqrt(n);

@ Resampling @
do while i<=b;
index=seqa(1,1,n);
@Set up re-sampling index@
rindex=submat(index,ceil(rndu(n,1)*n)',0);
x_b=x[rindex,.]; @Defines re-sample b
   vector of Xi*'s@

x_bar_b = meanc(x_b); @Calculating X-barb*
   for a re-sample@
x_bar_bt[i,1] = x_bar_b; @Setting up
   vector of
   bootstrapped x-bar*'s@

i=i+1; @Increase re-sampling index@

@Setting up for BCa zit calculation@
```

```
if x_bar_b <= x_bar;
prop=prop+1/b;
endif;

endo; @End of resampling loop@

@Constructing parametric confidence
interval @
pmlcl = x_bar - (tcal*sd_x_bar);
pmucl = x_bar + (tcal*sd_x_bar);

@Constructing normal
approximation confidence interval @
nalcl = x_bar - (tcal * stdc(x_bar_bt));
naucl = x_bar + (tcal * stdc(x_bar_bt));

@Constructing percentile confidence
interval @
x_bar_bt = x_bar_bt~sd_xb_bt; @Attach x-bar*'s
   and their sd's for sorting
   together (needed for percentile-t)@
x_bar_bt = sorthc(x_bar_bt,1); @Sort
   x_bar*'s vector @
@Defining the CI percentile
points for alpha = 0.05@
ll = round((B*0.05)/2);
ul = round((B-ll)+1);

@Defining lower and upper
percentile CI endpoints@
perlcl = x_bar_bt[ll,1];
perucl = x_bar_bt[ul,1];

@Constructing percentile-t confidence interval@
@Standardize the x-bar*'s using
the SD from each re-sample and the
   full sample mean@
t_boot = (x_bar_bt[.,1] - x_bar) ./ x_bar_bt[.,2];
t_boot = sorthc(t_boot,1); @Sort the
   vector of t_boot@

@Define upper and lower percentile-t endpoints.
NOTE: _Add_ t_boot conversion to both limits@
prtlcl = x_bar + (t_boot[ll] * sd_x_bar);
prtucl = x_bar + (t_boot[ul] * sd_x_bar);

@Constructing BCa confidence interval@

@Step #1: Calculate the acceleration
   factor, a @
@Jackknifing for BCa @
do while j<=n;
xj=submat(x,j,0); @Defines xj as vector with
   only case j of x @
xjack=setdif(x,xj,1); @Defines xjack as those
   case in x, but not in xj (so all but
   case j)@
xjck_bar = meanc(xjack); @Calculate the mean of
   xjack)@
mn_jack[j,1] = xjck_bar; @Placing the x-bar(i)
   into the jackknifed x-bar vector@
j=j+1; @Increasing jackknife index@
endo; @End jackknifing loop@

@Calculating acceleration factor, a,
   from the jackknifed
   vector of x-bar's@
mxjck = meanc(mn_jack); @Mean of jackknifed
   x-bar‘s@
a_num = sumc((mxjck-mn_jack)^3); @Numerator for
    acceleration factor@
@Calculate the denominator for a, setting
    format for
    small numbers@
a_den = ((C*(sumc((mxjck-mn_jack)^2))^1.5);
    format /rd 2,9;
if a_den == 0;  @To assure division
    by 0 doesn‘t occur@
a=0;
else;
a = a_num/a_den; @Calculate a @
endif;

@Step #2: Calculate z0 @
@Making sure prop =/ 0 or 1 (so zit =/ infinity)@
if prop == 0;
prop = .0001;
elseif
prop == 1;
prop = .9999;
else;
prop = prop;
endif;
zp=0;zit=-3.00; @Set starting points for
   z0 calculations@
do until zp > prop; @Loop to increment zit = z0
   until it is the z-score for
   the proportion of
   x-bar*'s below the full sample
   x-bar@
zp=cdfn(zit); @zp is the z-score associated
   with zit@
zit=zit+.001; @Increment zit @
endo;
zit = zit-0.001; @Delete the last zit increment @

@Step #3: Construct the BCa confidence interval@
@Define the lower and upper BCa .05
```

```
   percentile points@
bc_l_per =
cdfn(zit+((zit-1.9C)/(1-(a*(zit-1.9C)))));
bc_u_per =
cdfn(zit+((zit+1.9C)/(1-(a*(zit+1.9C)))));

@Calculate the bootstrap vector case number
for the lower and upper BCa bounds @
l_bca = round(B*bc_l_per); @Lower endpoint
   case number@
if l_bca == 0; @To avoid case #0
   being chosen@
l_bca = 1;
else;
l_bca = l_bca;
endif;
u_bca = round(B*bc_u_per); @Upper endpoint
   case number@
if u_bca > B; @To avoid case #(B+1)
   being chosen@
u_bca = B;
else;
u_bca = u_bca;
endif;
bcalcl = x_bar_bt[l_bca,1]; @Lower BCa
   CI endpoint@
bcaucl = x_bar_bt[u_bca,1]; @Upper BCa
   CI endpoint@
```

Textbook Review: *Estimation and Inference in Econometrics*, Russell Davidson and James MacKinnon, Oxford, 1993, 874 pp.

Nathaniel Beck
University of California, San Diego

Davidson and MacKinnon have pulled off a minor miracle: they have written an econometrics text that is new and different. An instructor trained with the classic Johnston text would have little difficulty using most new texts. Such is not the case with Davidson and MacKinnon. It is different, and one has to begin reading it at Chapter 1. Even the most jaded reader of TPM will find some new insights in Davidson and MacKinnon's approach. This, of course, means that we cannot teach with this text and continue to use our old notes and sequence of topics. My intuition is that the more standard, albeit excellent, texts, such as Greene, will therefore dominate our advanced methods courses. Meaning no insult to Greene, et al., this would be a shame.

Davidson and MacKinnon rely heavily on intuitive geometric concepts. Their excellent first chapter is entirely on the geometry of least squares. Rather than a complicated algebraic treatment, the argument is based on projecting on and off various linear subspaces. By the end of the first chapter, the student has a good understanding of the intuition of ordinary least squares, including the geometry of influential points.

The second chapter then dives immediately into the geometry (and some algebra) or non-linear least squares. While it may be disconcerting to some readers to treat OLS as simply a special case of NLLS, the approach is very intuitive. Chapter 2, like chapter 1, is heavily based on the logic of linear spaces, and is filled with pictures of regressors projected onto non-linear manifolds.

Estimation and asymptotic theory are taken up in the following several chapters. The treatment is remarkable: it is both intuitive and rigorous; the chapter on asymptotic theory is both lucid and correct. Their discussion of testing is first-rate. Unlike most texts, Davidson and MacKinnon give a substantial treatment of the direction of tests, and the implications of this for power against a variety of alternatives. There are also clear chapters on alternative methods of estimation, such as instrumental variables, maximum likelihood and generalized method of moments. While the book does not emphasize time series, it does have a good treatment of unit roots and co-integrations.

Davidson and MacKinnon have been important contributors to various econometric literatures. The book draws heavily on these contributions, and in particular, on the use of artificial regressions (Gauss-Newton iterations) for testing. This makes their treatment of both transformations and limited dependent variables more novel than what would be found in the standard texts. The book ends with a longish chapter on Monte Carlo methods. This chapter will serve as an excellent guide for those who wish to pursue this now popular technique.

In short, Davidson and MacKinnon is a fabulous combination of rigour and intuition, covering all the major topics while providing much new (at least for a text) material. It is the most exciting text I have seen. Having said that, it is simply too hard for our graduate students at UCSD (we use Gujarati), and it lacks the wonderful examples that make the Greene text so understandable. But even if you can't use this book as a text, get it for your library. But get it prepared to read it, and not to use as just another reference text.

Review of *The Phantom Respondents*, John Brehm, University of Michigan Press, 1993

Mark Fey
California Institute of Technology

The Michigan Studies in Political Analysis is a series devoted to the "innovative scholarship in the field of methodology in political science and the social sciences." The third volume in this excellent series, *The Phantom Respondents: Opinion Surveys and Political Representation* by John Brehm, capably meets this goal. It deals with the growing problem of survey nonresponse and the implications of this behavior on studies of public opinion and politics. It develops and tests a model of survey participation and offers a correction for the biases introduced by nonresponse. While dealing with fairly technical issues, the book is written in an accessible style, with welcome touches of humor sprinkled in with the multivariate probits.

The book opens with the argument that surveys and polls are more than just tools to measure public opinion, but rather are a modern vehicle for political representation and participation. Included in this argument is the useful reminder that the universe of polls and surveys consists of more than just the NES, GSS, and commercial polls that we are used to seeing, but also includes the vast number of surveys conducted by government agencies. Moreover, these latter surveys have a direct effect on public policy, such as the extension of unemployment insurance based on high unemployment figures from the CPS or the allocation of federal funds that are tied to counts of population, poverty, or joblessness. In sum, polls and surveys are an important part of scientific research, governmental policy-making and political representation. However, the author claims that each of these pursuits is vulnerable to the threat posed by survey nonresponse; to the extent that nonrespondents differ from respondents, the results of polls and surveys will be inaccurate and misleading.

As a first step in the study of nonresponse, Brehm poses the question "Who is missing?" To answer this question, Brehm compares the demographics of respondents to the NES and GSS from 1978 through 1988 and three NES telephone surveys from 1982, 1984 and 1988 with benchmark surveys, the Current Population Studies (CPS) for the same years. These comparisons lead to the conclusion that the academic surveys underestimate the number of young people, men, and wealthy families, and overestimate the proportion of the elderly, black respondents, the poor, and the less-educated. While such raw statistics are informative, they are of little help in assessing the impact of and correcting for nonresponse.

Chapter 5, "How Survey Nonresponse Damages Scientific Research" is of the most interest to political methodologists. It begins with an excellent introduction to the statistical consequences of nonresponse, illustrating the inconsistency of both univariate and multivariate estimates. The chapter then reviews Heckman's model of incidental selection and reports the results of Monte Carlo simulations on the effect that the covariance of the errors across the outcome and selection equations and the response rate have on estimation biases. Finally, the chapter discusses several procedures for modeling nonresponse and correcting these biased parameter estimates. The first approach, developed by Heckman (1976, 1979) and Achen (1986), is a two-stage method that involves estimating the likelihood of nonresponse and then incorporating this estimate into the outcome model in order to estimate the covariance between selection and outcome. The appropriate modification of this procedure for dichotomous dependent variables is also discussed. As an alternative, maximum likelihood methods of correcting for censored samples are mentioned in passing.

The "heart of [the] book," as Brehm calls it, is Chapter 6 which applies the two-stage correction to various empirical models in political science. The chapter reestimates models of turnout, vote choice, candidate evaluations, respondent's policy positions, and family income, but is ultimately disappointing. The correction for nonresponse, in the models that Brehm reestimates, does not change the parameter estimates very much, much less the sign of a coefficient of interest. This is not to say that nonresponse is of no concern to researchers — the potential exists for serious inaccuracy, especially with the apparent trend toward ever increasing rates of nonresponse, as laid out in Chapter 7. But the reader is left with the impression that for existing data sets, the only consistent change offered by the correction is in the constant term. This helps explain the well-known overreporting of turnout in NES samples. Brehm argues that since voters who are the most likely to turnout are also the most liking to respond to surveys, the surveys oversample citizens who vote. This is a pleasing result, but the constant term is usually of little interest. For the coefficients with substantive meaning, the correction has less of an effect, changing the magnitude of the estimates to varying degrees. While it is of course true that more accurate parameter estimates are intrinsically valuable, the author neglects to compare the bias introduced by nonresponse to the bias caused by other data problems common to political methodologists such as question wording and ordering, model misspecification, missing data, and measurement error. The reader is left with considerable uncertainty as to the relative importance of nonresponse among the class of methodological data problems.

Overall, *The Phantom Respondents* is a worthwhile read that deserves special attention by those interested and engaged in quantitative analysis of political science. We as consumers of, and occasionally producers of, survey research need to be aware of the importance of and issues surrounding survey nonresponse. By making the dangers of nonresponse clear and offering helpful tools to deal with them, this book performs a valuable service to the field.

Call for Proposals for the Twelfth Annual Political Methodology Summer Conference

Larry Bartels
Princeton University

The Twelfth Annual Political Methodology Summer Conference, organized by the Methodology Section of the American Political Science Association, will be held July 27-30, 1995, at Indiana University, Bloomington, Indiana. Interested faculty and advanced graduate students are invited to apply to attend.

Attendance will be limited to approximately 35 participants, with roughly equal representation of senior faculty, junior faculty, and graduate students. The aim of the conference is to spur the development, evaluation, and application of quantitative and qualitative research methods in political science. Previous conferences have typically consisted of six three-hour sessions, each featuring two papers and two discussants, plus additional informal sessions to discuss graduate student research.

Faculty applicants should submit a curriculum vitae and a brief (3-5 page) outline of research they propose to present. Original research in all areas of political methodology is welcome. Papers may propose new research methods, evaluate existing methods, or apply existing methods to substantive problems of interest to political scientists. Methodological work with important implications for comparative politics, international relations, formal theory, or public policy is especially welcome.

Graduate student applicants should submit a curriculum vitae and a letter describing their methodological interests. Graduate student applicants should also arrange for a faculty member to submit a letter of recommendation including a brief description of the student's methodological training and research program. Graduate students who are invited to attend will be asked to prepare in advance of the meeting a brief (3-5 page) statement of their major research interests.

Some conference expenses for faculty participants and all expenses for graduate student participants (including travel to and from Bloomington, food, and lodging) will be defrayed by the Methodology Section (through a grant from the National Science Foundation) and by Indiana University.

Applications to attend the conference must be received by March 31, 1995. Applications will be reviewed by a committee appointed by the president of the Methodology Section, and invitations to attend the conference will be extended by May 1, 1995. For further information or to apply, contact Larry M. Bartels, Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, NJ 08544-1013. E-mail: lbartels@wws.princeton.edu.

Call for Nominations for Best Paper Award

Bill Berry
Florida State University

Next year, the Political Methodology section of APSA will inaugurate an annual award for the best conference paper on a political methodology topic presented during the previous year. Eligible papers include those presented at the annual summer methodology meeting, the APSA annual meeting, or any of the regional (e.g., Midwest, Southern, Western) conferences. The first award will be presented at the 1995 APSA Annual Meeting in Chicago.

The committee seeks nominations from section members for the award. Members are urged to send as many as three nominations to any of the committee members by no later than January 31, 1995. For each paper nominated, please provide the author(s), the title and the name of the 1994 meeting (Wisconsin summer meeting, NY APSA meeting, or any regional political science meeting) at which the paper was presented.
Committee members can be reached as follows:

Bill Berry
Department of Political Science
Florida State University
Tallahassee, FL 32306-2049
FAX: 904-644-6041; voice: 904-644-7321
wberry@garnet.acns.fsu.edu

Henry Brady
Department of Political Science
210 Barrows Hall
University of California, Berkeley 94720
phone: 510-642-3008
hbrady@bravo.berkeley.edu

Doug Rivers
Department of Political Science
Stanford University
Stanford, CA 94305
rdr@titus.stanford.edu

The Political Methodologist

Maib de
alioia Dtitte o Techolog
PaDdea

The Political Methodologist is the newsletter of the Political Methodology Section of the American Political Science Association. We gratefully acknowledge the support of the California Institute of Technology in helping to defray the editorial and production costs of the newsletter, and Abby Delman for her assistance.

Subscriptions to *TPM* are free to members of the APSA's Methodology Section. Please contact APSA to join the section. Dues are $8.00 per year.

Submissions to *TPM* are welcome. Articles should be sent to the co-editors if possible, by e-mail to rma@hss.caltech.edu and beck@ucsd.edu. Alternatively, submissions can be made on diskette as plain ascii files sent to R. Michael Alvarez, Division of Humanities and Social Sciences 228-77, California Institute of Technology, Pasadena, CA 91125. LaTeX format files are especially encouraged. The deadline for submissions for the next issue is January 1, 1995.