Microeconometrics and Stata over the Past Thirty Years

A. Colin Cameron
Department of Economics, University of California - Davis.

This version: August 28, 2014

Abstract

This article discusses how microeconometrics research has evolved since 1985, the year Stata was released, and how Stata has been part of this process.

JEL Classification: C10, C21, C23, C87.
Keywords: microeconometrics; Stata; history.

In preparation for a special issue of the Stata Journal.

Contents

I. Introduction

Microeconomics research has become much more empirical oriented over the past thirty years. This has been made possible by much greater computational power. The IBM XT 286, introduced in 1986, had 640 KB of RAM, a 6 MHz processor, a 20MB hard disk and a 1.2 MB processor (Wikipedia 2014). By contrast a typical PC today runs more than 500 times faster with memory and storage that is more than 10,000 times larger. This greater computer power has been accompanied by increased data availability, new methods, and the development of statistical software to implement these methods.

In this paper I summarize how theoretical and applied microeconometrics research has evolved over the past thirty years and how Stata has been part of this process. The discussion of theory is necessarily brief, with further detail provided in Cameron (2009). The role of Stata, one of several packages available to econometricians, is especially important as it is now the most commonly used package in applied microeconometrics.

The interplay between theory and implementation is not straightforward as considerable time can pass from the introduction of new methods to their use by applied researchers and their incorporation in a statistical package. This delay is partly due to it taking time before the usefulness of the new method is clear. To some extent this is a chicken and egg problem, as methods are used much more once they are incorporated into a statistical package. And delay also arises because some methods, notably semiparametric regression, maximum simulated likelihood and Bayesian methods, are difficult to code into a user-friendly command that will work for a wide range of problems. Because Stata is programmable this process can be, and has been, speeded up by users developing their own code ahead of any official Stata command. In some cases this code is made available to other Stata users as a user-written Stata ado file. This short article mentions only a few of these useful add-ons.

II. Regression and Stata

Many of the core regression methods now widely used in applied microeconometrics research were introduced in the late 1970's and early 1980's. These methods include sample selection models (Heckman 1976), quantile regression (Koenker and Bassett 1978), bootstrap (Efron 1979), heteroskedastic robust standard errors (White 1980, 1982) and generalized method of moments (GMM) estimation (Hansen 1982). Additionally several seminal books appeared in the early and mid 1980's, namely Maddala (1983) for limited dependent variable models, Amemiya (1985) for nonlinear regression models, and Hsiao (1986) for panel data.

Cox (2005) provides a brief history of the first twenty years of Stata; Baum, Schaffer and Stillman (2011) provide a recent overview. Stata was introduced in 1985 for use on IBM PC's running under DOS, rather than on a mainframe computer. The initial release of Stata

was quite limited, focusing on tools for data management and exploratory data analysis, due to both its newness and the low computing power of PC's. The only regression command in the initial release was command `regress` for least squares estimation of linear models.

The basic limited dependent variable models were among the first regression models to be introduced into Stata – logit and probit (1987), survival models (1988), Tobit models and multinomial logit models (1992), and linear sample selection models and negative binomial models (1993). Quantile regression methods became much more widely used after their incorporation in Stata in 1992. Commands for general nonlinear least squares and maximum likelihood estimation were introduced in 1993. GMM estimation was incorporated in several linear model commands, though a general command for GMM estimation was not introduced until 2009. The basic panel data commands, a strength of Stata, were introduced in 1995 (linear) and 1996 (nonlinear).

Increased computing power has enabled greater use of simulation methods. Monte Carlo experiments based on a known data generating process can be conducted in Stata using command `simulate` or command `postfile`. Random variables can be drawn directly from a wide range of distributions following a major Stata enhancement in 2008. These distributions including the multivariate normal and the truncated multivariate normal (using the GHK simulator). The Stata random number generators include Halton and Hammersly sequences in addition to a standard random uniform generator.

Methods for simulation-based estimation of parametric models were developed in the 1980's and 1990's, especially maximum simulated likelihood (MSL) estimation (McFadden 1989, Pakes and Pollard 1989) and Bayesian Markov chain Monte Carlo (MCMC) methods (Geman and Geman 1984). These methods have enabled the estimation of increasingly complex parametric models. In empirical microeconometrics these are most often limited dependent variable models such as the random parameters logit model. Furthermore, Bayesian methods are generally used merely as a tool; the results are still given a frequentist interpretation rather than a Bayesian interpretation.

It is difficult to provide robust general code for these methods. Stata instead uses these methods in particular contexts, notably in command `asmprobit` that estimates the multinomial probit model using MSL and in multiple imputation commands that use MCMC methods. Additionally, user-written code provides Stata front ends to the Bayesian statistical packages Winbugs (Thompson, Palmer and Moreno 2006) and MLwiN (Leckie and Charlton 2013).

Where possible Stata avoids use of simulation-based estimation methods. In particular, complex parametric models are often difficult to estimate due to an intractable integral. For a one-dimensional integral, such as that in the linear random effects model, it is standard to use Gaussian quadrature rather than simulation methods. For higher dimensional integrals

of the multivariate normal that appear in mixed models, Stata commands `mixed` and `gsem` use adaptive multivariate Gaussian quadrature rather than simulation methods.

An alternative strand of research has developed methods to estimate regression models that rely on relatively weak distributional assumptions. The building block is nonparametric regression on a single regressor. Several methods have been proposed in the statistics literature, beginning with kernel regression in 1964, followed by Lowess, local polynomial regression, wavelets and splines. Stata initially provided Lowess estimation. Local polynomial regression, including kernel regression and local linear as special cases, appeared as command `lpoly` in 2007. These nonparametric regression commands, and the kernel density estimation command `kdensity`, are especially valuable for viewing data and key statistical output such as residuals.

The single-regressor nonparametric regression methods do not extend well to models with more than one regressor, due to the curse of dimensionality. Econometricians have been at the forefront of developing semiparametric models that combine a high-dimensional parametric component with a low-dimensional (usually single-dimensional) component. The late 1980's and early 1900's saw development of estimation methods for three commonly-used models – the partial linear model, the single-index model, and generalized additive models. Semiparametric methods are particularly useful for limited dependent variable models with censoring and truncation as they enable crucial parametric assumptions on unobservables to be weakened; Pagan and Ullah (1999) provide a survey. These semiparametric methods generally require selection of smoothing parameters, sometimes with deliberate undersmoothing or oversmoothing. Perhaps for this reason there are no official Stata commands for semiparametric regression, though there are some Stata add-ons for some specific estimators. The lack of semiparametric regression commands in Stata is one reason that semiparametrics methods, a focus of recent theoretical econometrics research, are infrequently used in applied microeconometrics.

In addition to obtaining regression coefficients under minimal assumptions, the econometrics literature has developed methods for statistical inference under minimal assumptions. Heteroskedastic-robust standard errors were developed by White (1980, 1982) and introduced into Stata in ?? If model errors are clustered, then default and heteroskedastic-robust standard errors can be much too small. Extensions to cluster-robust inference were made by Liang and Zeger (1986) and Arellano (1989). The early inclusion of cluster-robust standard errors (Rogers 1993) in basic Stata regression commands greatly increased their usage. Even though Stata is at the forefront in providing robust standard errors, however, their incorporation into more advanced estimation commands has taken considerable time.

When standard errors, non-robust or robust, are not available they can be obtained by an appropriate bootstrap. A bootstrap command appeared in Stata in 1991 with significant

enhancement in 2003. The theoretical literature has emphasized a second use of the bootstrap, namely bootstraps with asymptotic refinement that may lead to better finite-sample inference. These latter bootstraps are seldom used in practice; a notable exception is the wild cluster bootstrap when there are few clusters (Cameron, Gelbach and Miller 2008). Bootstraps with refinement can be implemented in Stata – bias-corrected confidence intervals as a bootstrap option and other methods with some additional coding.

A distinguishing feature of econometrics is the desire to make causal inference from observational data. Instrumental variables estimation and its extension to GMM were the dominant methods when Stata was introduced. Papers by Nelson and Startz (1990) and Bound, Jaeger and Baker (1995) highlighted the need for alternative inference methods when instruments are weak. Recent results on weak instrument asymptotics for linear models with non-i.i.d. model errors, the usual case in empirical microeconomics studies, are implemented in Stata add-ons `ivreg2` (Baum, Schaffer and Stillman 2007) and `weakiv` (Finlay, Magnusson and Schaffer 2013).

A major change in causal microeconometrics research is use of the potential outcomes framework of Rubin (1974) that has evolved into the quasi-experimental or treatment effects literature, summarized in Angrist and Pischke 2010. For selection on observables only one can use matching methods such as propensity score matching (Rosenbaum and Rubin 1983), or use inverse-probability weighting. A Stata module to implement these methods, introduced in 2013, superseded earlier user-written add-ons. When selection is also on unobservables most methods can be implemented using existing Stata commands. These methods include local average treatment effects (LATE) estimation (Imbens and Angrist 1994), a reinterpretation of IV when treatment effects are heterogeneous, fixed effects panel models and their extension to differences in differences with repeated cross-section data, sample selection models, and regression discontinuity design. For dynamic linear panel models with fixed effects the methods of Arellano and Bond (1991) and extensions can be implemented using the official Stata command `xtabond` and the user-written add-on `xtabond2` (Roodman 2003).

Methods for spatially correlated data have been progressively developed over the past thirty years. At this stage there are no official Stata commands for spatial regression, but there are several user-written Stata add-ons that handle and analyze spatial data, including the spatial regression module `SPPACK` (Drukker, Peng, Prucha and Raciborski 2011).

Researchers in biostatistics and in social sciences other than economics, who are also Stata users, employ some regression methods that are not often used in empirical microeconometrics. Generalized linear models (command `glm`) and generalized estimating equations (`xtgee`) cover a range of nonlinear regression models including those with binary or count dependent variable. Mixed models or hierarchical models (command `mixed`) can lead to more precise estimation when model errors are clustered than if a simple random effects model is esti-

mated. And other social sciences make greater use of completely specified structural models (commands `sem` and `gsem`).

III. Empirical Research and Stata

There is more to empirical research than obtaining parameter estimates and their standard errors, the subject of the previous section.

The first step is to simply analyze and view the data ahead of any regression analysis. Useful graphical methods are kernel density estimates and two-way scatter plots with a fitted nonparametric regression curve. Stata introduced a very rich publication-quality graphics package in 2003. Interpreting the sources of variation in grouped data is simplified by using the `statsby` command and `xt` commands such as `xtsum`, `xttab` and `xtdescribe`.

Model diagnostics and specification tests can be useful. Applied microeconometrics studies tend not to use available methods that can detect outlying observations and influential observations. This is in part due to concerns about subsequently overfitting a model, though such diagnostics can also highlight mistakes such as miscoded data. Available model specification tests are infrequently used, notable exceptions being Hausman tests and tests of overidentifying restrictions. Stata post-estimation commands include these standard methods as well as enabling in-sample and out-of-sample prediction.

Many applied studies in microeconometrics seek to estimate a marginal effect, such as the increase in earnings with one more year of schooling, rather than a regression model parameter per se. Marginal effects, and their associated standard errors, can be computed using the `margins` command introduced in 2009 that supplanted the user-written command `margeff` (Bartus 2004). Factor variables, also introduced in 2009, enable extension to models with interacted regressors.

Empirical microeconomics studies are increasingly based on data sources that are very complex. Complications include: (1) data may come from several different sources; (2) data may come from surveys; (3) data may have a grouped structure such as panel data or individual-level data from several villages; and (4) some data may be missing.

A real contribution of Stata has been its ability to handle these complications. Stata is a data management package, in addition to a statistical package, with features including ability to handle string variables and commands to merge and append datasets. The Stata survey commands control for weighting, clustering and stratification. Empirical microeconometrics studies generally do not use the survey commands. Instead regular estimation commands are used with weights, if necessary, and with appropriate cluster-robust standard errors. Stratification is ignored, with some potential loss in estimator efficiency. Grouped data can be manipulated using the `by` prefix commands and the `reshape` command. Stata's estimation

commands automatically allow for missing data using case deletion. If case-deletion is not valid then one can use weighted regression, if weights are available. Alternatively one can use the Stata multiple imputation module introduced in 2009. For imputation empirical economics researchers currently rely on case-deletion, or on crude imputation methods such as hot deck imputation, despite their limitations.

Stata was initially limited in the size of dataset it could handle as it requires that all data be stored in memory, in order to speed up computations. This limitation has greatly diminished over time given increases in computer memory capacity and the emergence of 64-bit PCs.

As empirical studies have become more complex, the need for replicability has increased. Researchers need to be able to keep track of their own work, return to it after leaving it for a considerable period of time, and potentially coordinate computations with coauthors, research assistants and students. Furthermore, several leading journals require that data and programs be posted at their archives. Stata is well-suited to producing replicable studies as it is command driven, and the resulting Stata scripts can be run on a wide range of platforms and on newer versions of Stata.

As is clear from the previous section, it can take considerable time before a new method is included in a statistical package such as Stata. It is therefore advantageous to use software that is programmable. Stata has always been programmable, and includes a complete matrix programming language Mata that was introduced in 2007.

The widespread use of Stata has had the advantage of creating a community of users. Stata encourages this community through the Stata Technical Bulletin (began in 1990 and superseded by the Stata Journal in 2001), Statalist Server (1994), Stata users Group meetings (1995), the Stata website (1996), and Stata Press books (1999). For basic applied microeconometrics the books by Baum (2006), Cameron and Trivedi (2010) and Mitchell (2012) are especially helpful. The websites for introductory econometrics texts provide code for analysis in Stata. The Statistical Software Components (1997) website provides many Stata user-written programs that can be directly downloaded to Stata. As already noted, Stata users have provided many useful add-on programs. While some have been superseded by official Stata commands, many still fill gaps or augment official Stata commands.

As is the case for any statistical package, the ubiquity of Stata also has downsides. Data analyses may be restricted only to what is easily implemented in Stata. Researchers may not understand the limitations of the methods used, such as Tobit model estimates relying on very strong parametric assumptions. And at some stage Stata may become legacy software, yet one with a very large user base. To date Stata has avoided this by continuing to target academic researchers in economics, other social sciences, and biostatistics.

IV. References

Amemiya, T. 1985. *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Angrist, J.D., and J.-S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Arellano, M. 1987. Computing robust standard errors for within-group estimators. *Oxford Bulletin of Economics and Statistics* 49: 431-434.

Arellano, M., and S. Bond. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* 58: 277–298.

Bartus, T. 2005. Estimation of marginal effects using margeff. *The Stata Journal* 5(3): 309-329.

Baum, C.F. 2006. *An Introduction to Modern Econometrics using Stata*. College Station, TX: Stata Press.

Baum, C.F., M.E. Schaffer, and S. Stillman. 2011. Using STATA for applied research: Reviewing its capabilities. *Journal of Economic Surveys* 25(2): 380-394.

Baum, Christopher F., Mark E. Schaffer, Steven Stillman. 2007. Enhanced routines for instrumental variables/GMM Estimation and Testing. The Stata Journal 7(4): 465-506.

Bound, J., D.A. Jaeger, and R.M. Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90: 443-450.

Cameron, A.C. 2009. Some recent developments in microeconometrics. In *Palgrave Handbook of Econometrics, Volume 2: Applied Econometrics*, ed. T.C. Mills and K. Patterson, 729-774. Palgrave Macmillan, London.

Cameron, A.C., J. Gelbach, and D.L. Miller. 2008. Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics* 90: 414-427.

Cameron, A.C., and P.K. Trivedi. 2010. *Microeconometrics using Stata*. Revised ed. College Station, TX: Stata Press.

Cox, N.J. 2005. A brief history of Stata on its 20th anniversary. *The Stata Journal*, 5(1): 1-18.

Drukker, D.M., H. Peng, I. Prucha, and R. Raciborski. 2011. SPPACK: Stata module for cross-section spatial-autoregressive models. Statistical Software Components S457245, Boston College Department of Economics.

Efron, B. 1979. Bootstrapping methods: another look at the jackknife. *Annals of Statistics*, 7, 1-26.

Finlay, K., L. Magnusson, and M.E. Schaffer. 2013. WEAKIV: Stata module to perform weak-instrument-robust tests and confidence intervals for instrumental-variable (IV) estimation of linear, probit and tobit models. Statistical Software Components S457684, Boston

College Department of Economics.

Geman, S., and D. Geman. 1984. Stochastic relaxation, Gibbs distributions and Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6: 721-741.

Hansen, L.P. 1982. Large sample properties of generalized methods of moments estimators. *Econometrica* 50: 1029-1054.

Heckman, J.J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5: 475-492.

Hsiao, C. 1986. *Analysis of Panel Data.* Cambridge, UK: Cambridge University Press.

Imbens, G.W., and J. Angrist. 1994. Identification and estimation of local average treatment effect. *Econometrica* 62: 467-475.

Koenker, R., and G. Bassett. 1978. Regression quantiles. *Econometrica* 46: 33-50.

Leckie, G. and C. Charlton. 2013. runmlwin - A program to run the MLwiN multilevel modelling software from within Stata. *Journal of Statistical Software* 52(11): 1-40.

Liang, K.-Y., and S.L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73: 13-22.

Maddala, G.S. 1983. *Limited-Dependent and Qualitative Variables in Economics.* Cambridge, UK: Cambridge University Press.

McFadden, D. 1989. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica* 57: 995-1026.

Mitchell, M.N. 2012. *A Visual Guide to Stata Graphics.* 3rd ed. College Station, TX: Stata Press.

Nelson, C.R., and R. Startz. 1990. The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *Journal of Business* 63: S125-140.

Pagan, A.R., and A. Ullah. 1999. *Nonparametric Econometrics.* Cambridge, UK: Cambridge University Press.

Pakes, A.S., and D. Pollard. 1989. Simulation and the asymptotics of optimization estimators. *Econometrica* 57: 1027-1057.

Rogers, William H. 1993. Regression standard errors in clustered samples. *Stata Technical Bulletin* 13: 19-23.

Roodman, D. 2009 How to do xtabond2: An introduction to difference and system GMM in Stata. *The Stata Journal*, 9(1): 86-136.

Rosenbaum, P. and D.B. Rubin. 1983. The central role of propensity score in observational studies for causal effects. *Biometrika* 70: 41-55.

Rubin, D. 1974. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* 66: 688-701.

Thompson, J.R., Palmer, T.M., and S. Moreno. 2006. Bayesian analysis in Stata using WinBUGS. *The Stata Journal* 6(4): 530-549.

White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48: 817-838.

White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50: 1-25.

Wikipedia. 2014. http://en.wikipedia.org/wiki/IBM_Personal_Computer_XT, accessed August 26, 2014.