The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from Sage Publishing via telephone 805-499-9774 (U.S. customers) or 44-(0)20-73248701 (International), email journals@sagepub.com, or online at

https://journals.sagepub.com/home/stj

The *Stata Journal* (ISSN 1536-867X) is a quarterly journal of Stata Press and is published by Sage Publishing in association with StataCorp LLC. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LLC.

The Stata Journal (2023)
23, Number 4, pp. 1062–1073 DOI: 10.1177/1536867X231212446

Review of A. Colin Cameron and Pravin K. Trivedi's Microeconometrics Using Stata, Second Edition

Sebastian Kripfganz
University of Exeter Business School
Exeter, U.K.
S.Kripfganz@exeter.ac.uk

Abstract. In this article, I review *Microeconometrics Using Stata, Second Edition*, by A. Colin Cameron and Pravin K. Trivedi (2022, Stata Press).

Keywords: gn0097, book review, regression, cross-sectional data, panel data, nonlinear models, causal inference, Stata

1 Introduction

The (revised) first edition of Cameron and Trivedi's *Microeconometrics Using Stata* quickly became a standard reference on the syllabi for undergraduate and graduate courses, and it proved very popular among applied researchers in economics and the broader social sciences. The second edition carries an ever higher weight, literally—it comes as two volumes, with 1,675 pages combined. It is not just a revised version brought up to date with recent methodological developments and enhancements to Stata; it now has a much broader scope and covers several topics not included in the first edition.

Volume I of the *Second Edition* touches various aspects of applied microeconometric work at an introductory level, with a focus on the linear regression model. It also presents some fundamentals for nonlinear regression models. *Volume II* provides an in-depth discussion of the latter and some more advanced topics: duration models, spatial regression, machine learning, and Bayesian methods, among others. While the two volumes can be bought separately, they should be seen as one entity. Chapters in both volumes are frequently cross referenced, and almost everyone should find interesting and relevant content in both. Generally speaking, earlier chapters present basic introductions, and later chapters present more advanced discussions.

Not least because of its volume, this *Second Edition* is unlikely to be a book one would read from beginning to end in one go. In teaching, some chapters will inevitably be skipped, depending on the level and scope of the course. Practitioners of applied microeconometrics—the target audience—might selectively study the chapters relevant for their work. One characteristic of this book is that chapters do not hierarchically build on each other. One can just open to any chapter of interest. Yet readers might find it useful to go back and forth between different chapters on related topics.

A. Colin Cameron, distinguished professor at the University of California–Davis, and Pravin K. Trivedi, distinguished professor emeritus at Indiana University–Bloomington and honorary professor at the University of Queensland, draw upon extensive teaching and research experience in microeconometrics. Instead of just presenting established methods, they make a commendable effort to incorporate very recent developments. Often, it is a long journey for new theoretical results to find their way into the empirical practice, and it can take even longer for prevailing approaches to disappear when they are no longer deemed fit for purpose. Even if some of the new insights are mentioned only briefly with references for further reading, this new edition will help speed up the transmission of knowledge.

Many of the more recent advances are not yet implemented in official Stata commands. This is understandable, given the fluidity of frontier research and StataCorp's desire to provide a reliable product for empirical work—based on methods that have passed scientific scrutiny. Yet this gap is filled by the vast number of community-contributed commands. Despite the rise of other programming environments for statistical analyses, there does not appear to be a break in the upward trend of new contributions by and for the Stata community. Readers of *Microeconometrics Using Stata* will find many references to and examples making use of community-contributed commands, although naturally their coverage cannot be comprehensive.

It is not just the presentation of community contributions that separates the book from the official Stata reference manuals. The manual entries are primarily concerned with the particular capabilities of a specific command—often, though, providing quite extensive methodological details and examples. By contrast, the strength of *Microeconometrics Using Stata* is its focus on practical aspects that arise when answering certain empirical questions or when dealing with specific types of data and the connections they draw between different methods and models.

Near the end of each chapter, a section listing additional resources for further reading can be found. This includes links to the Stata manuals, community-contributed commands, and research articles. Furthermore, references to other textbooks are listed here, where readers can find more thorough coverage of the respective topics. One of them is the authors' own *Microeconometrics: Methods and Applications* (2005), which was closely aligned with the first edition of *Microeconometrics Using Stata*. The second edition has now grown out of the footsteps of its big brother and has developed its own identity.

However, the book should still be seen as a complement to rather than a substitute for traditional econometrics textbooks. While *Microeconometrics Using Stata* provides some background information on econometric theory, the reader is expected to be familiar with econometric notation and fundamental concepts. Even for adept researchers, the exposition of some material might be insufficient to develop a deeper understanding of the methods presented. The book should rather be seen as a starting point in the quest of expanding one's knowledge. It provides excellent intuition, with hands-on Stata examples using both real-world and simulated data.

The book often only scratches the surface of a specific Stata command's capabilities. Researchers typically have needs going beyond the examples shown, depending on the nature of their application or the idiosyncrasies of their data. Thus, even after one studies a chapter of *Microeconometrics Using Stata*, consulting the Stata manuals or help files for full documentation and additional examples remains inevitable, as well as further reading in theory-oriented textbooks and original research articles.

In the final section of each chapter, the readers are invited to improve their familiarity with the Stata commands presented by solving a (relatively small) number of practical exercises. The datasets used throughout the two volumes are available on a companion website, which not only facilitates learning by doing but also provides a valuable resource for university lecturers.

2 Volume I

Volume I kicks off with a quick start for new users on how to navigate around Stata. The first chapter is mostly a compressed version of the official *Stata User's Guide*. Readers with prior Stata experience can safely skip this and also much of the following chapter.

Chapter 2 provides an introduction to different data formats and storage types, and it alerts the reader to some important pitfalls arising from numerical precision problems or from inputting data in an inappropriate format. The initial stage of a research project often involves substantial data management tasks. The book walks the reader through a real-world example—from inputting data, over labeling and transforming variables, dealing with missing values, to selecting the relevant sample. This chapter briefly presents essential commands for dataset manipulation, followed by an introduction to frequently used graphs for data visualization. Given the scope of the book, this chapter can only glimpse at these topics. For those interested in wider coverage, Michael N. Mitchell's books *Data Management Using Stata: A Practical Handbook, Second Edition* (2020) and *A Visual Guide to Stata Graphics, Fourth Edition* (2022) deserve consideration.

Chapter 3 illustrates the different steps of linear regression analysis—from coefficient estimation, establishing statistical significance and interpreting coefficient estimates, testing simple hypotheses, presenting the results from several regressions with the new `etable` and `collect` commands (introduced in Stata 17), to specification analysis and tests. Instead of directly jumping to the model estimation, the reader is reminded to carry out basic data checks and consider useful variable transformations. Throughout the book, the authors emphasize the importance of robust standard errors, which should in most cases be the researcher's default choice, even though they are not the Stata default. Consequently, heteroskedasticity-robust and cluster–robust standard errors are introduced at an early stage, followed by an intuitive introduction to the bootstrap approach. The book is more than just a handbook on how to run a series of commands: it gives lots of practical advice. Even experienced researchers might find useful reminders in this chapter. For example, the authors emphasize that specification tests have power in more than one direction; if a test rejects the null hypothesis of homoskedasticity, this could be a consequence of other model misspecifications, such as the conditional mean.

Chapter 4 covers several aspects of model predictions, a more detailed treatment of marginal effects, and an introduction to regression decomposition analysis and difference-in-differences estimation. The discussion of the `margins` command is quite comprehensive and demonstrates its power even for the linear regression model.

Not all chapters address the same target group. Many practitioners might be happy to just apply standard tools for estimation and inference. For them, the book is helpful in developing a structured approach to microeconometric analysis and obtaining guidance on the respective methods. Others might want to go further and develop a deeper understanding of a method's statistical properties; they will find the book a helpful resource too. Chapter 5 introduces Monte Carlo simulations as a verification tool for statistical methods or as a means of assessment for their robustness to deviations from standard settings. That this chapter appears so early in the book reflects the authors' belief that Monte Carlo experiments are underutilized as a powerful pedagogical tool.

Chapter 6 covers approaches for clustered data, from cluster–robust standard errors to generalized least-squares estimation and systems of linear regressions. It is especially insightful to learn how random-effects and fixed-effects panel-data methods (`xtreg`) can be adapted for clustered cross-sectional data. The book also deserves credit for introducing mixed models with heterogeneous slopes, which are still often treated as orphans by microeconometricians.

Linear instrumental-variables estimation is comprehensively introduced in chapter 7, including a brief illustration of the `etregress` command for fitting treatment effects models with a binary endogenous regressor and a latent-variable first stage. A large part of this chapter is devoted to the topical issue of weak instruments, including critical remarks on the still widely used rule of thumb that the first-stage F statistic should exceed 10, which recent research no longer deems appropriate. In this context, community-contributed commands (`weakivtest` [Pflueger and Wang 2015], `ivreg2` [Baum, Schaffer, and Stillman 2003], `condivreg` [Mikusheva and Poi 2006], `rivtest` [Finlay and Magnusson 2009], and `weakiv` [Finlay, Magnusson, and Schaffer 2014]) are indispensable, given that most of the latest developments are not yet accessible through official commands. Although this is still an active area of research and some of the methods might be supplanted again in the future, the benefits of presenting these new methods certainly outweigh this risk.

In chapter 8, the reader finds an extensive treatment of the `xtreg` command and the related methods for linear panel-data methods. Aside from the Stata examples, this chapter does not add much compared with standard textbooks on this topic. On rare occasions, here and elsewhere, there are typos and inaccuracies. For example, even two of them appear on page 414, where the authors first refer to "T fixed effects", while there are N of them. Shortly thereafter, the authors argue that weak or sequential exogeneity—which would allow for a lagged dependent variable—satisfies the requirements of the first-difference estimator, but this is factually incorrect.[1] Both of these

1. In chapter 9, it is correctly stated that the first-difference estimator is inconsistent when lagged dependent variables are introduced.

inaccuracies were already present in *Microeconometrics Using Stata, First Edition*. Yet such mistakes are pardonable given the extent of the book.

Chapter 9 presents panel-data instrumental-variables estimators, from Hausman–Taylor for static models to Arellano–Bond and system generalized method of moments (GMM) estimators for dynamic models, as well as methods for long panels. While the topics in chapters 8 and 9 are slightly rearranged compared with the first edition, the content remains largely unchanged. There is no discussion of mixed models in the panel-data context anymore (which are sufficiently covered elsewhere in this book), while some coverage has been added of interactive effects and heterogeneous panels. A few more recent developments in this area are missing, especially on long panels. Admittedly, though, the latter are not typically seen in microeconometric applications. Overall, the chapters on panel-data models are not the strong suit of the book, but they do cover the necessary fundamentals and provide a useful starting point for further study.

Chapter 10 returns to cross-sectional data and introduces nonlinear regression models, with `probit` and `logit` as the leading cases for binary outcomes. This chapter briefly mentions nonlinear least-squares estimation and the generalized linear model, but detailed coverage is postponed until later.

In chapter 11, the authors note recent discussions in the scientific community about the downsides of focusing on statistical significance (rather than economic significance) and the correct use of p-values. Besides the correct interpretation of p-values, there is growing concern about distorted inference due to pretesting and selective reporting of results. The authors also direct the reader's attention to power calculations for desired effect sizes. After they introduce general testing principles and relevant distributions, illustrations follow of linear and nonlinear Wald tests, likelihood-ratio tests, and Lagrange multiplier tests. Important additions to the *Second Edition* are discussions of controlling the familywise error rate or the false discovery rate for multiple testing and the `power onemean` command for power calculations. A minor point of criticism relates to the presentation of one-sided Wald tests. For the alternative hypothesis $H_a\colon \beta_j > 0$, the authors consider the complementary null hypothesis $H_0\colon \beta_j \leq 0$, which is a set of all parameter values on the nonpositive real line. It is then not entirely accurate to say that $z = \widehat{\beta}_j / s_{\widehat{\beta}_j} \overset{a}{\sim} N(0,1)$ "under H_0". The distribution of this supremum statistic is obtained only at the boundary of H_0, where $\beta_j = 0$, not for the whole set $\beta_j \leq 0$. Interestingly, there is no consensus among econometrics textbook authors on whether or not to formulate the null hypothesis of a one-sided test as a set hypothesis $H_0\colon \beta_j \leq 0$ or a point hypothesis $H_0 : \beta_j = 0$—both versions exist. The latter avoids the above complication by imposing the implicit assumption that the parameter space is restricted to nonnegative values. For practical considerations, this distinction hardly matters.[2]

Chapter 12 demonstrates bootstrap methods using the `vce(bootstrap)` option, the `estat bootstrap` postestimation command, or the `bootstrap` prefix command. A nice illustration is given of how to implement a bootstrap program for the Hausman test when we do not want to assume that one of the estimators is fully efficient. A new section in the *Second Edition* covers the wild bootstrap, implemented in the popular

2. I thank Samuel Engle for an insightful conversation on this topic.

community-contributed `boottest` package (Roodman et al. 2019), which is especially useful for clustered data with few clusters and for instrumental-variables estimation with heteroskedastic errors and weak instruments. The chapter further illustrates how to implement alternative resampling schemes using the `bsample` command and concludes with a brief description of the jackknife method.

The estimation of nonlinear regression models is the topic of chapter 13. Given the authors' own research interests, it is not surprising that a count-data application serves as the main illustration. This chapter helps one develop a basic understanding of maximum likelihood (ML) estimation, nonlinear least squares, generalized linear models (which are more popular in areas of applied statistics other than economics), and the GMM. The authors reemphasize the importance of robust (or cluster–robust) standard errors: a whole section is dedicated to different variance–covariance estimators and another one to clustered data. The latter illustrates again that panel-data commands can be employed even without actual panel data. The demonstration of how to implement two-step estimators with GMM by stacking moment conditions can be very useful for various applications. The chapter furthermore includes a more detailed treatment of marginal effects, which naturally play an important role in the context of nonlinear models.

Chapter 14 walks the reader through the estimation of finite-mixture regression models, polynomial regressions, and spline and piecewise regressions before embarking on an introductory tour through nonparametric regression, with a focus on kernel-weighted polynomial regression for one explanatory variable. Separating the introductory remarks in this chapter (and other chapters in *Volume I*) from the more comprehensive discussion in later chapters in *Volume II* might seem slightly confusing, but it allows readers new to this area to have easier access to the material without becoming overwhelmed by some of the more advanced content.

In chapter 15, the reader finds an introduction to quantile regression. It starts with the usual motivation of conditional quantiles based on minimizing the absolute prediction error loss. Asymmetric penalties for overprediction or underprediction give rise to different conditional quantiles as the optimal predictor. Aside from this foundation, quantile regression is more robust to outliers and provides a more complete picture of the statistical relationship between the variables of interest rather than just looking at the conditional mean. The authors also highlight that quantile regression provides a way of accommodating heterogeneous treatment effects. It is useful to be reminded about the difference between the conditional and the unconditional distribution and the consequence for interpreting estimates at different quantiles.

Volume I concludes with appendixes on programming in Stata and Mata, including the Mata optimization functions `moptimize()` and `optimize()`. While nice to have, the value added of these appendixes is debatable, given that only limited programming knowledge is required elsewhere in the book. It would be preferable if some of this content was embedded in the following chapter 16. Would-be programmers might want to go straight to Christopher F. Baum's book *An Introduction to Stata Programming, Second Edition* (2016) and William W. Gould's *The Mata Book: A Book for Serious Programmers and Those Who Want to Be* (2018).

3 Volume II

Volume II largely covers nonlinear regression models. Its first chapter focuses on optimization methods. While such optimization typically occurs under the hood of an estimation command, some knowledge of these methods is still useful to understand where estimation might go wrong. Moreover, if no command exists for a particular model, users might have to implement an ML or GMM estimator themselves. The chapter includes discussion of Stata's `ml` command and Mata's `moptimize()` and `optimize()` functions and some hints on program debugging. As a minor criticism, in addition to the example presented using `optimize()` with evaluator `d2`, it would have been instructive to demonstrate GMM estimation using `moptimize()` with evaluator `q1`.

Chapter 17 provides a more in-depth discussion of binary-outcome models (logit, probit, complementary log–log) and some remarks on fractional outcomes. Readers find the necessary reminder to use `vce(robust)` if the functional form might be misspecified or if the linear probability model is used because the latter's regression errors are necessarily heteroskedastic. A few alternative models are presented, such as the heteroskedastic probit model or the generalized logit model, as well as nonparametric logit estimation for prediction purposes. A useful discussion on endogeneity in binary outcome models can be found here as well. Besides putting different approaches into context, the book is especially useful in illustrating ways to implement certain techniques for which no dedicated command exists. Here, although only very briefly, this applies to nonlinear instrumental-variables estimation with the `gmm` command.

Chapter 18, on multinomial models, is updated to accommodate the `cm` commands for choice models, which were introduced in Stata 16 and replaced the earlier `asclogit` and `asmprobit` commands for (alternative-specific) conditional logit and multinomial probit models, respectively. Nested logit models are discussed for the estimation of additive random-utility models, which relax the "independence of irrelevant alternatives" assumption. In the context of multinomial probit models, a brief introduction is provided for the maximum simulated likelihood method. The chapter concludes with discussion of ordered outcomes and some remarks on multivariate outcomes.

The focus of chapter 19 is on the tobit model for censored or truncated data. The authors emphasize that consistency of the estimator depends on the possibly strong normality assumption, which should be checked using model diagnostics. The demonstration of conditional moment tests for normality and homoskedasticity with generalized

residuals from the tobit estimation is especially valuable, given that no official Stata package for them exists. A useful trick is shown for setting the censoring point with lognormal data or when the true censoring point is unknown. Two-part or hurdle and sample-selection models can be used to overcome some of the limitations of the tobit approach, as explained subsequently in this chapter. To sidestep the reliance of traditional parametric selection models on functional-form and distributional assumptions, the authors present copula models as a flexible alternative, using the community-contributed `heckmancopula` command (Hasebe 2013). A discussion of missing values and panel attrition follows, which also covers inverse-probability weighting and endogenous sample selection.

Chapter 20 praises the robustness benefits of the Poisson estimator for count data (and more), requiring only the correct specification of the conditional mean. Robust standard errors are promoted to allow for overdispersion. As an alternative, the negative binomial model is presented. For the hurdle model as an approach to dealing with excess zeros, the authors explain in detail how to implement the two-step procedure. Finite mixture models, which take a significant portion of this chapter, account for richer forms of unobserved heterogeneity, with the zero-inflated count model as a special case. The chapter then discusses the three leading approaches for dealing with endogeneity in nonlinear models: ML estimation of a structural model, control function estimation, and nonlinear instrumental-variables or GMM estimation. A useful discussion, although without examples, is given on how to proceed with clustered data. Finally, this chapter demonstrates how to apply quantile regression to count data with the community-contributed `qcount` command (Miranda 2007), which uses jittering to smooth the count data.

Readers already familiar with the first or revised edition will find a new topic in chapter 21, which introduces methods for survival analysis. This includes an illustration of the Cox proportional hazards model as a semiparametric approach, as well as a discussion of parametric duration models. The different implications of the varying approaches for the hazard function are illustrated insightfully.

In nonlinear panel-data models, individual effects pose additional complications. Chapter 22 addresses these issues, with the usual distinction between fixed effects and random effects (primarily in the way economists use this terminology). Often, fixed-effects estimators are unavailable because of the incidental-parameters bias and the inability to eliminate the individual effects. The bias-correction approach to the estimation of nonlinear fixed-effects models is still underutilized in empirical practice. While this chapter introduces the community-contributed `logitfe` and `probitfe` commands (Cruz-Gonzalez, Fernández-Val, and Weidner 2017), they would deserve a more extended treatment. Their key benefit is the provision of bias-corrected estimates of marginal effects, and they are applicable to dynamic models as well. Another attractive approach illustrated here is "correlated random-effects" estimation.

The remaining chapters cover further topics that were not included in the previous edition. Chapter 23 follows a different approach than the previous chapters in that it takes a command-based perspective and presents the methods largely in the con-

text of specific examples. The `fmm` prefix command is presented for finite-mixture or latent-class models with unobserved heterogeneity before the `me` commands for nonlinear mixed-effects models move into focus. For handling endogenous regressors, the chapter then looks at linear and generalized structural equation models, which can be fit using `sem` and `gsem`. While these commands are very flexible, they come with complicated syntaxes. The book makes some effort to explain the different syntax components and provides useful tables with `sem` or `gsem` syntax equivalent to other standard estimation commands, which can serve as a starting point to developing more complicated models. The structural equation model path builder is briefly mentioned but without detailed discussion. An instructive example for the use of latent variables is provided in the context of classical measurement error. The chapter concludes with an introduction to extended regression models, useful for dealing with endogeneity and nonrandom selection.

The next two chapters are concerned with treatment effects. Chapter 24 starts with an introduction of the potential-outcomes framework and randomized control trials, including a discussion of optimal sample-size and power analysis. The newly developed design-based inference approach is briefly mentioned but not explored further. Subsequently, treatment evaluation with observational data comes into focus when selection is on observables, with attention restricted to binary treatments. Methods discussed are regression adjustment, inverse-probability weighting, their doubly robust augmentations, and matching methods. Useful remarks are made about obtaining robust standard errors for the average treatment effect by stacking moment equations. A detailed example is presented using the Oregon Health Insurance Experiment. A discussion of multilevel treatments and conditional quantile treatment effects rounds off this chapter.

Chapter 25 provides an advanced introduction to parametric models for endogenous treatment using extended regression models and Stata's `et` commands, as well as quasi-experimental methods for heterogeneous treatment effects—which make use of several community-contributed commands—such as local average treatment effects, difference in differences, synthetic control, and regression discontinuity designs. This includes mention of some very recent difference-in-differences developments. The chapter concludes again with a look at quantile regression methods.

Spatial regression models are considered in chapter 26, which are primarily motivated for geospatial data but can also be used in a peer-effects context and with network data. The book unfortunately does not provide examples for the latter, aside from an exercise left to the reader. Stata's `sp` commands are illustrated with an example using geospatial data. The authors emphasize that thinking about the structure of the spatial-weights matrix is an important modeling step because misspecified spatial weights might turn the coefficient estimation inconsistent.

Although titled "Semiparametric regression", chapter 27 also takes a closer look at nonparametric kernel and series regression. Semiparametric approaches considered here include the partial linear model, the single-index model, and generalized additive models—estimable with the community-contributed commands `semipar` (Verardi and Debarsy 2012), `sls` (Barker 2014), and `gam` (Royston and Ambler 1998), respectively.

Nonparametric regression methods generally suffer from a curse of dimensionality when there are many (potential) regressors. For prediction purposes, machine learning techniques can be alternatives. These are introduced in chapter 28, which begins with a presentation of cross-validation and penalty measures to assess the (out-of-sample) predictive ability of the models. This is followed by an introduction to shrinkage estimation (ridge regression, lasso, elastic net), principal components, neural networks, regression trees, and random forests. Typical use cases are the partial linear model with many potential controls and instrumental-variables estimation with many potential instruments. Community-contributed commands `vselect` (Lindsey and Sheather 2010) and `gvselect` (Lindsey and Sheather 2015) are presented for best-subsets selection and stepwise selection. Subsequently, principal components is introduced as a dimension reduction technique. As opposed to such unsupervised learning, methods for supervised learning account for the outcome variable. Making use of several community-contributed commands, advanced machine learning methods—neural networks, regression trees, bagging, random forests, boosting, and classification and cluster analysis—are presented in very condensed forms. Yet this introduction is intuitive enough to demystify these techniques for readers new to machine learning or "big data". The book then discusses inference after using variable selection techniques, with a focus on the partialing-out lasso estimator. As the authors note, (causal) inference for machine learning methods "is an exceptionally active area of current econometric research, and we anticipate an explosion of new methods that will be implementable in Stata using one's own coding as community-contributed Stata programs and, ultimately in some cases, as official Stata programs" (p. 1522).

Toward the end of the book, the authors take the reader on a journey to the Bayesian universe. Chapter 29, the penultimate chapter, illustrates the `bayes` prefix command and the more flexible `bayesmh` command. The authors highlight that Bayesian methods are not necessarily incompatible with classical and frequentist thinking: with a noninformative prior and a focus on the posterior mode, Bayesian methods can be computational tools to obtain ML estimates, especially when the sample size is large. The authors also emphasize that the posterior results can be quite dependent on the prior specification for not just parameters of interest but also seemingly innocuous parameters.

The final chapter takes a closer look at the Markov chain Monte Carlo method. It provides helpful Mata code examples of the Metropolis–Hastings algorithm and the Gibbs sampler, which can be used as a starting point for one's own implementation. The chapter concludes with a treatment of multiple imputation methods, which can be regarded as an application of Bayesian methods.

4 Conclusion

Stata users nowadays have access to a wide range of learning resources. This includes the *Stata User's Guide* and the PDF reference manuals—often underappreciated sources of information—as well as help files, *Stata Journal* articles, online video tutorials, and countless other web resources. Despite this free competition, Cameron and Trivedi's

Microeconometrics Using Stata, Second Edition proves to be a valuable reference book for applied researchers. While highlights emphasized in this book review, as well as critical remarks made, are necessarily subjective, there should be something for everyone on the many pages of the two volumes.

5 Acknowledgment

This book review was written without help from artificial intelligence.

6 References

Barker, M. 2014. sls—Semiparametric least squares from Ichimura, 1993. GitHub. https://github.com/michaelbarker/stata-sls.

Baum, C. F. 2016. *An Introduction to Stata Programming*. 2nd ed. College Station, TX: Stata Press.

Baum, C. F., M. E. Schaffer, and S. Stillman. 2003. Instrumental variables and GMM: Estimation and testing. *Stata Journal* 3: 1–31. https://doi.org/10.1177/1536867X0300300101.

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

———. 2022. *Microeconometrics Using Stata*. 2nd ed. College Station, TX: Stata Press.

Cruz-Gonzalez, M., I. Fernández-Val, and M. Weidner. 2017. Bias corrections for probit and logit models with two-way fixed effects. *Stata Journal* 17: 517–545. https://doi.org/10.1177/1536867X1701700301.

Finlay, K., and L. M. Magnusson. 2009. Implementing weak-instrument robust tests for a general class of instrumental-variables models. *Stata Journal* 9: 398–421. https://doi.org/10.1177/1536867X0900900304.

Finlay, K., L. M. Magnusson, and M. E. Schaffer. 2014. weakiv: Stata module to perform weak-instrument-robust tests and confidencce intervals for instrumental-variable (IV) estimation of linear, probit and tobit models. Statistical Software Components S457684, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s457684.html.

Gould, W. 2018. *The Mata Book: A Book for Serious Programmers and Those Who Want to Be*. College Station, TX: Stata Press.

Hasebe, T. 2013. Copula-based maximum-likelihood estimation of sample-selection models. *Stata Journal* 13: 547–573. https://doi.org/10.1177/1536867X1301300307.

Lindsey, C., and S. Sheather. 2010. Variable selection in linear regression. *Stata Journal* 10: 650–669. https://doi.org/10.1177/1536867X1101000407.

———. 2015. Best subsets variable selection in nonnormal regression models. *Stata Journal* 15: 1046–1059. https://doi.org/10.1177/1536867X1501500406.

Mikusheva, A., and B. P. Poi. 2006. Tests and confidence sets with correct size when instruments are potentially weak. *Stata Journal* 6: 335–347. https://doi.org/10.1177/1536867X0600600303.

Miranda, A. 2007. qcount: program to fit quantile regression models for count data. Statistical Software Components S456714, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s456714.html.

Mitchell, M. N. 2020. *Data Management Using Stata: A Practical Handbook*. 2nd ed. College Station, TX: Stata Press.

———. 2022. *A Visual Guide to Stata Graphics*. 4th ed. College Station, TX: Stata Press.

Pflueger, C. E., and S. Wang. 2015. A robust test for weak instruments in Stata. *Stata Journal* 15: 216–225. https://doi.org/10.1177/1536867X1501500113.

Roodman, D., M. Ø. Nielsen, J. G. MacKinnon, and M. D. Webb. 2019. Fast and wild: Bootstrap inference in Stata using boottest. *Stata Journal* 19: 4–60. https://doi.org/10.1177/1536867X19830877.

Royston, P., and G. Ambler. 1998. sg79: Generalized additive models. *Stata Technical Bulletin* 42: 38–43. Reprinted in *Stata Technical Bulletin Reprints*. Vol. 7, pp. 217–224. College Station, TX: Stata Press.

Verardi, V., and N. Debarsy. 2012. Robinson's square root of N consistent semiparametric regression estimator in Stata. *Stata Journal* 12: 726–735. https://doi.org/10.1177/1536867X1201200411.

About the author

Sebastian Kripfganz is a senior lecturer in econometrics at the University of Exeter Business School, where he teaches econometric methods at the undergraduate and graduate level. He is the author of several methodological research papers and Stata estimation commands.