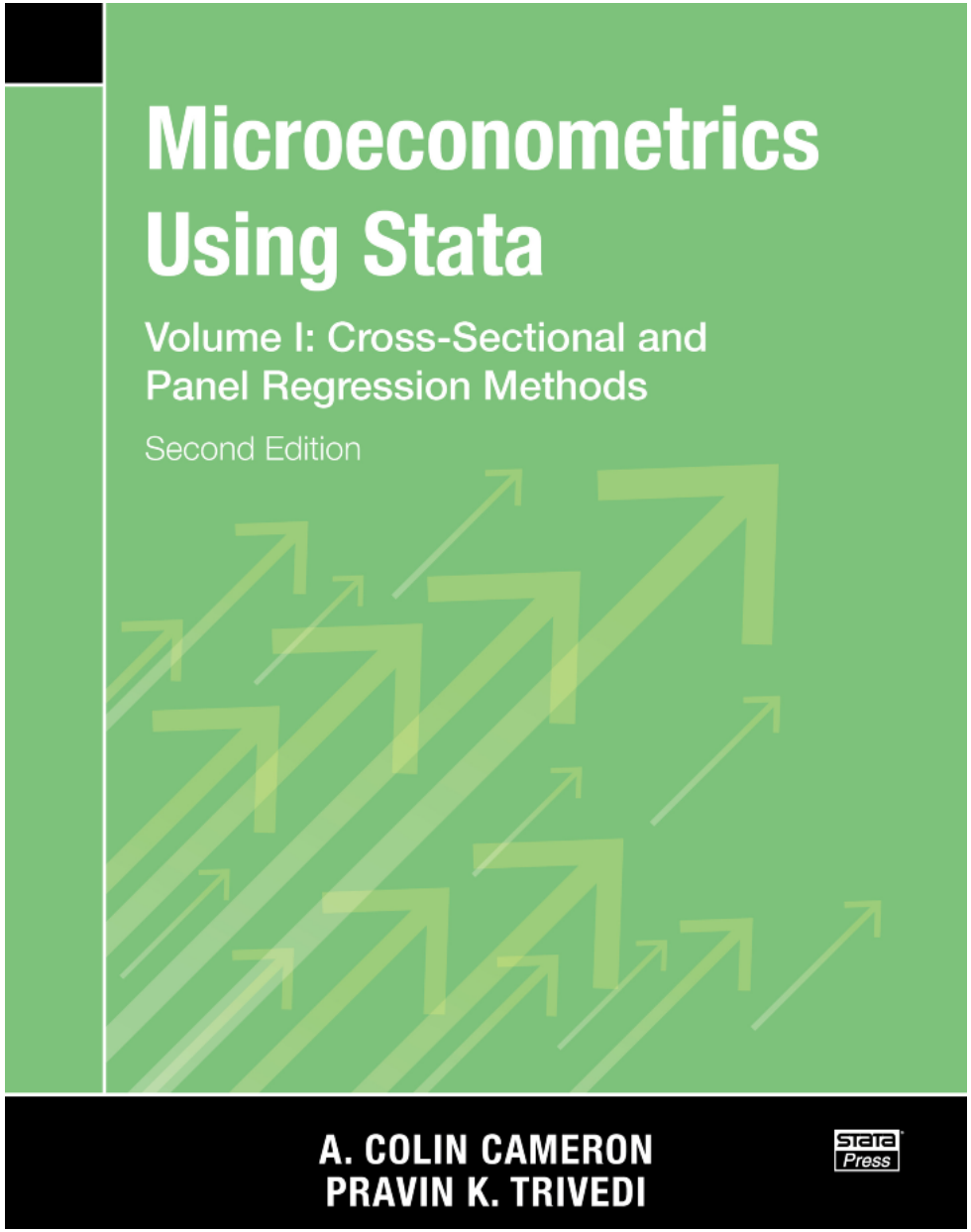
What did I learn by writing the Second Edition of Microeconometrics Using Stata? (Published July 2022 by Stata Press)

A. Colin Cameron (coauthor Pravin Trivedi)
Univ. of California Davis

February 27, 2023

Talk Outline

1. Book Chapters
2. My Coauthor
3. Why did we write the Stata book?
4. What did I learn about Stata?
5. What did I learn about modern econometrics?
6. Will I make money?

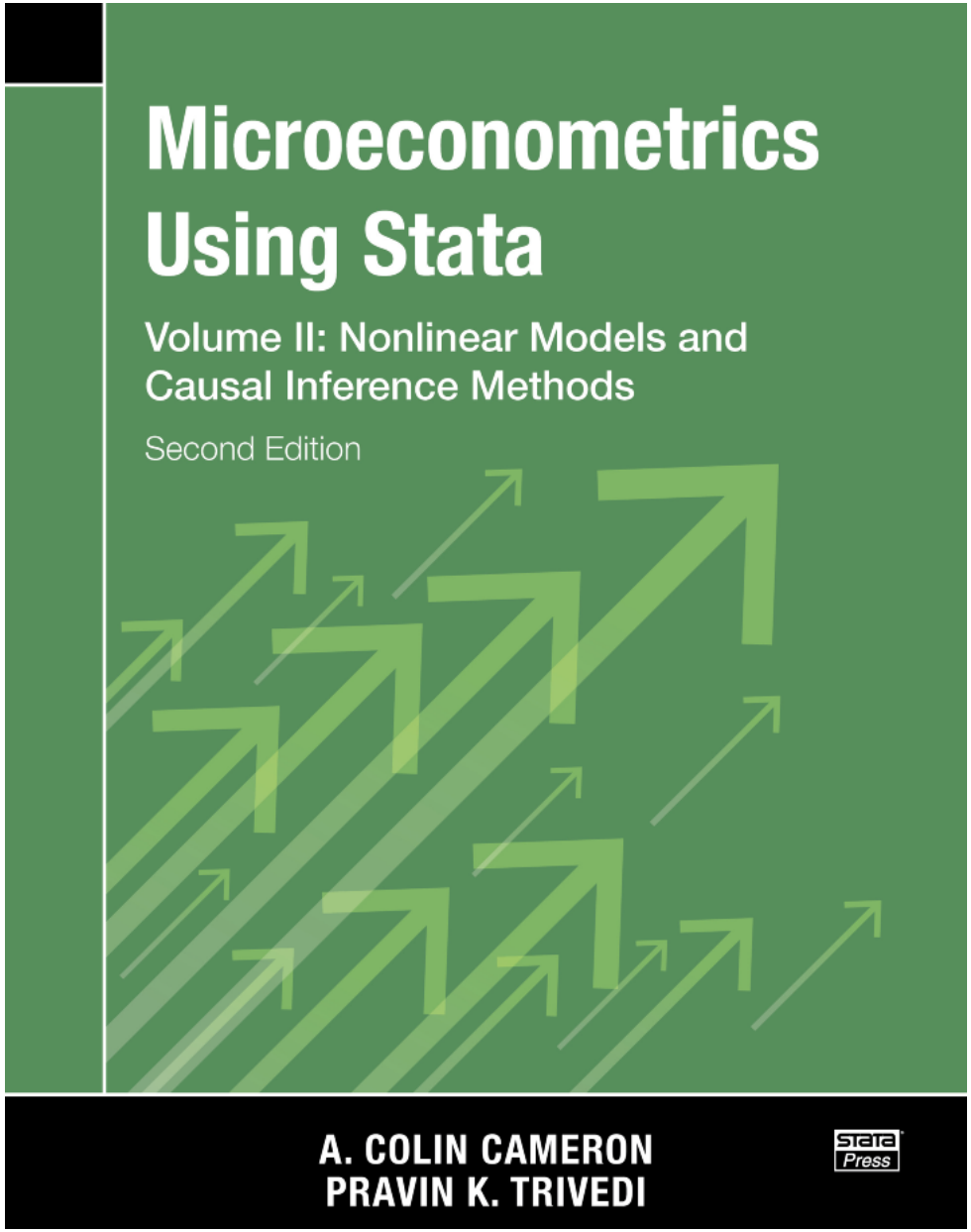

Microeconometrics
Using Stata
Volume I: Cross-Sectional and
Panel Regression Methods
Second Edition
A. COLIN CAMERON
PRAVIN K. TRIVEDI
Stata Press

1. Stata basics
2. Data management and graphics
3. Linear regression basics
4. Linear regression extensions
5. Simulation
6. Linear regression with correlated errors
7. Linear instrumental variables regression
8. Linear panel-data models: Basics
9. Linear panel-data models: Extensions
10. Introduction to nonlinear regression
11. Tests of hypotheses and model selection
12. Bootstrap methods
13. Nonlinear regression methods
14. Flexible regression: finite mixtures and nonparametric
15. Quantile regression

Appx.A: Programming in Stata

Appx.B: Mata

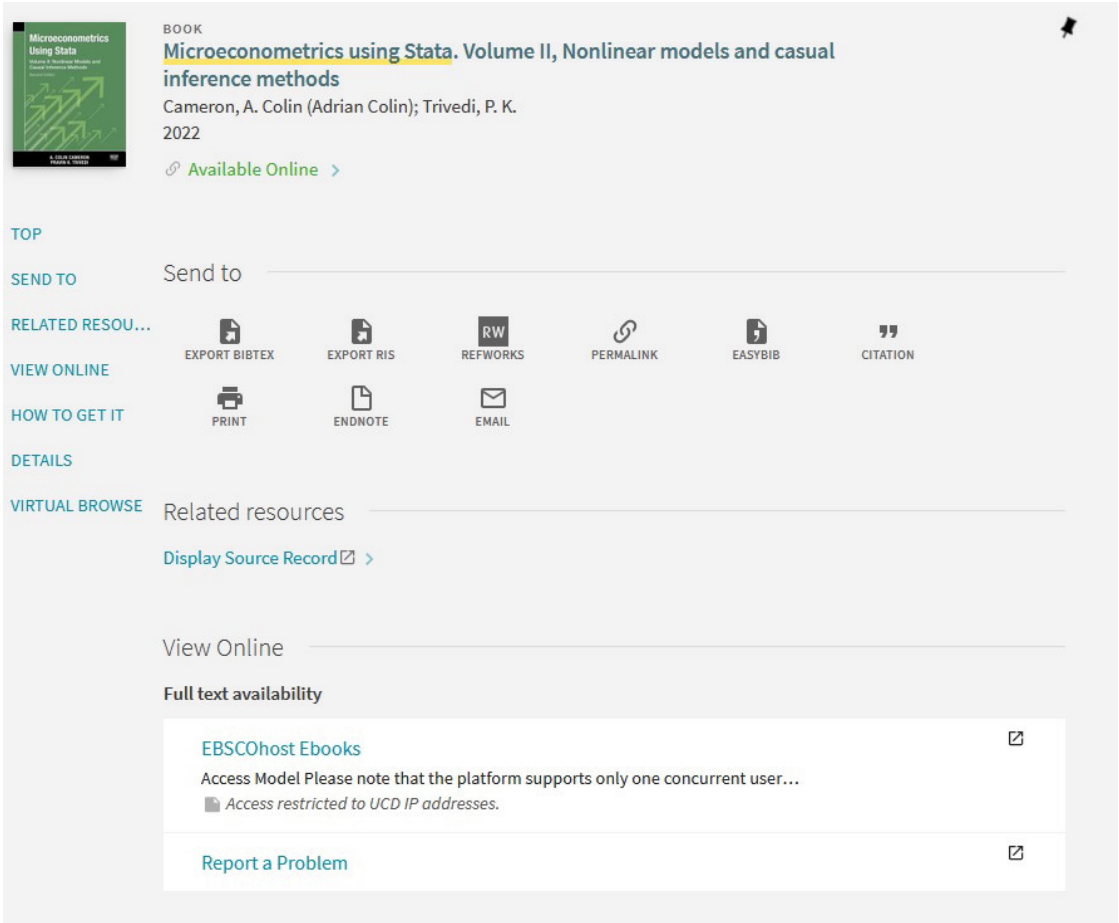
Appx.C: Optimization in Mata


Microeconometrics
Using Stata
Volume II: Nonlinear Models and
Causal Inference Methods
Second Edition
A. COLIN CAMERON
PRAVIN K. TRIVEDI
Stata Press

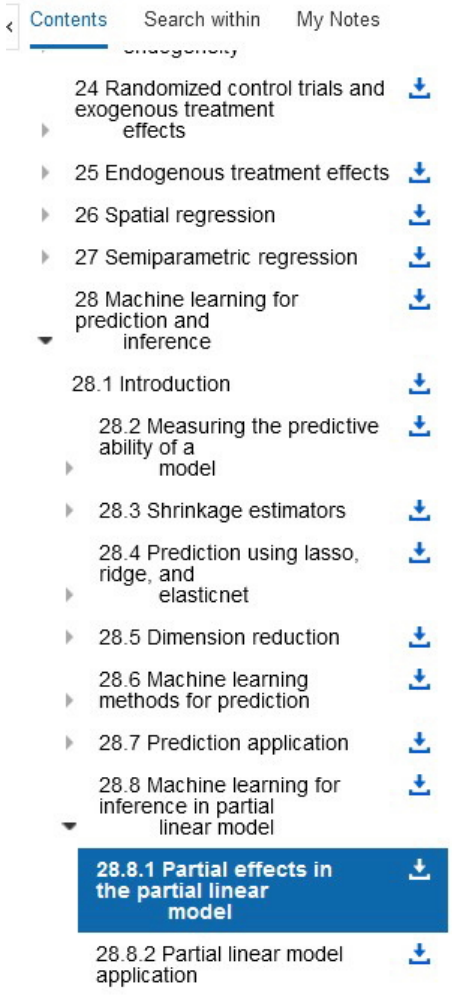
16. Nonlinear optimization methods
17. Binary outcome models
18. Multinomial models
19. Tobit and selection models
20. Count-data models
21. Survival analysis for duration data
22. Nonlinear panel models
23. Parametric models for heterogeneity and endogeneity
24. RCTs and exogenous treatment effects
25. Endogenous treatment effects
26. Spatial regression
27. Semiparametric regression
28. Machine learning for prediction and inference
29. Bayesian methods: Basics
30. Bayesian methods: MCMC algorithms

Digital Version available through UCD Library / Ebsco

- Not all universities will provide such access.



Example Section (+ can download up to 100 pages)



in version 16.

28.8.1 Partial effects in the partial linear model

We consider a setting where interest lies in measuring the partial effect on y of a change in variables $\mathbf{d}$, controlling for additional control variables.

A partial linear model for linear regression specifies

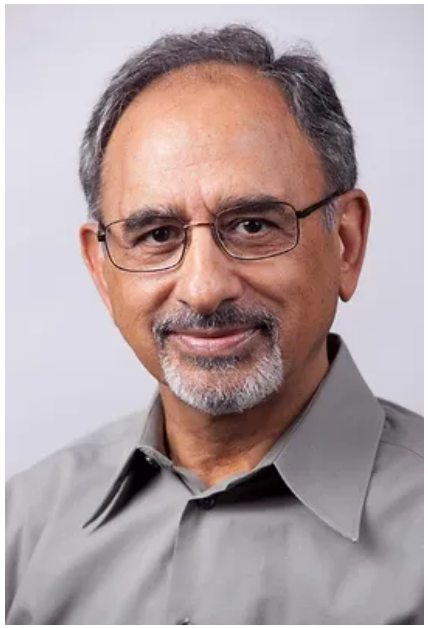
$$y = \mathbf{d}'\boldsymbol{\alpha} + g(\mathbf{x}_c) + u$$

where $\mathbf{x}_c$ denotes selected control variables and $g(\cdot)$ is a flexible function of $\mathbf{x}_c$. The parameter α can be given a causal interpretation with the selection-on-observables-only assumption that $E(u|\mathbf{d}, \mathbf{x}_c) = 0$. The goal is to obtain a root-N consistent and asymptotically normal estimator of the partial effect α.

The partial linear model was introduced in section 27.6. There $g(\cdot)$ was unspecified, and estimation was by semiparametric methods that required that there be few controls $\mathbf{x}_c$ to

2. My Coauthor

- Pravin Trivedi (Ph.D. LSE 1970) is Distinguished Professor Emeritus and J. H. Rudy Professor Emeritus in the Department of Economics at Indiana University - Bloomington.
- I took several courses from him as an undergraduate at The Australian National University.
- I returned to Australia in my fourth year of Ph.D. to work as research assistant on a project with him and two others on the link between health services demand and health insurance.



3. Why did we write the book?

- 1. "Regression Analysis of Count Data" research monograph published 1998
  - proud to post code on the web, using Limdep
  - so people asked "do you have Stata code?"
  - hence use Stata.
- 2. "Microeconometrics: Methods and Applications" published 2005
  - goal was to provide an accessible graduate level text
    - ★ for advanced empiricists (not theorists)
    - ★ for those who don't have high level course work
    - ★ for those with gaps (e.g. not see nonparametrics)
  - did not know of Wooldridge (2002) until it was published
  - limited applications with Stata 8 code posted on web.

Why did we write the book (continued)?

- 3. "Microeconometrics using Stata" published 2009, 2010
  - follow-up to MMA with a lot more Stata code
  - initial plan was to again publish with Cambridge University Press
  - instead switched to Stata Press
  - good decision as code (in Stata 10.1 and 11) is cleaner.
- 4. "Microeconometrics using Stata" 2nd edition published 2023
  - update both methods (closer to frontier) and Stata 17
  - took seven years!!

4. What did I learn About Stata?

- For econometrics Stata is reasonably close to frontier
  - mostly cross-section and short panel - less time series.
  - (matrix) programmable so allows element-by-element operators
  - Mata is very distinct from Stata
    - ★ so learning Mata is like learning Matlab or Python
  - can go back and forth between Stata and Python.

What did I learn About Stata (continued)?

- From Version 11 to 17 Stata has added many things, including
  - factor variables
  - test power computations
  - new random number generator (so needed to redo a lot of output)
  - treatment effects
  - richer nonparametric methods
  - richer multinomial choice models
  - machine learning (lasso and ridge)
  - finite mixture models
  - structural equation models (linear and nonlinear)
  - multilevel mixed effects (nonlinear hierarchical models)
  - Bayesian methods and multiple imputation.

What did I learn About Stata (continued)?

- Stata covers not just econometrics methods.
  - produce output for documents with improved tables command
  - dynamic linking of Stata output to document using dyndoc
  - Stata graphics are quite powerful - I use scripts
  - data frames allow several datasets
    - ★ supplants preserve and restore
  - biostatistics users now bigger than econometrics
  - many courses, conferences and webinars
  - Stata is becoming more corporate.

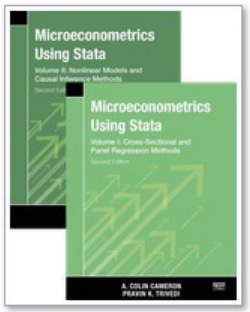

Microeconometrics
Using Stata
Volume II: Nonlinear Models and
Causal Inference Methods
Second Edition
COLIN CAMERON
PRAVIN K. TRIVEDI
Microeconometrics
Using Stata
Volume I: Cross-Sectional and
Panel Regression Methods
Second Edition
STATA

What did I learn About Stata (continued)?

- Data, programs ($>$ 10,000 lines) and index are available free at https://www.stata-press.com/books/microeconometrics-stata/

Microeconometrics Using Stata, Second Edition

Volume I: Cross-Sectional and Panel Regression Methods
Volume II: Nonlinear Models and Causal Inference Methods



Click to enlarge

Inside preview

Print | eBook | Kindle

$169.00 Print set

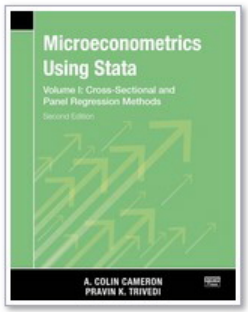
Buy now

New edition

Authors: A. Colin Cameron and Pravin K. Trivedi
Publisher: Stata Press
Copyright: 2022
ISBN-13: 978-1-59718-359-8
Pages: 1,675; paperback
Price: $169.00

Preface to the Second Edition
Download the datasets used in this book (from www.stata-press.com)

Chinese and Korean translations available of previous edition



Click to enlarge

Volume I: Cross-Sectional and Panel Regression Methods

Print | eBook | Kindle

$109.00 Print

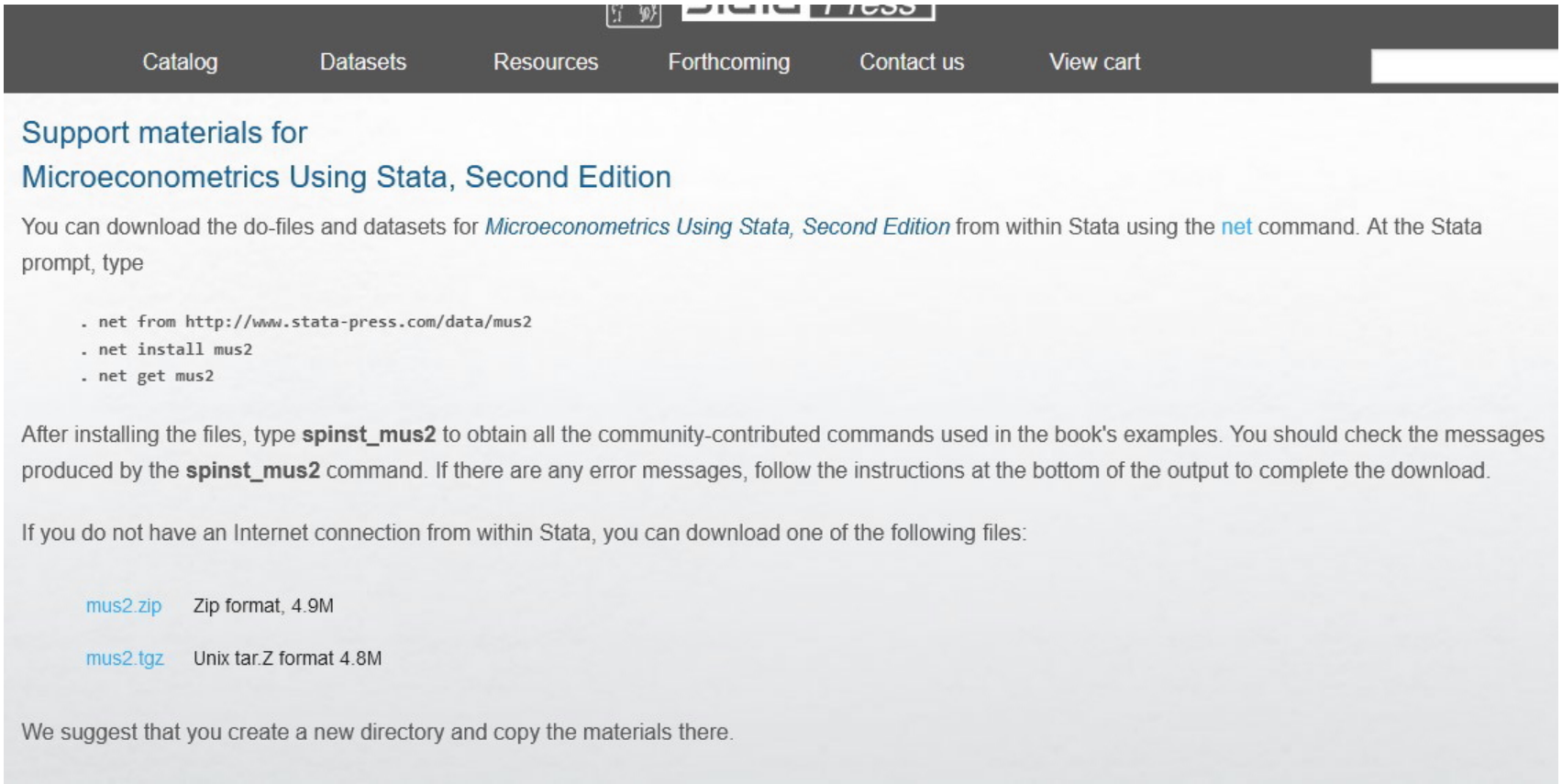
Buy now

ISBN-13: 978-1-59718-361-1
Pages: 817; paperback
Price: $109.00

Author index
Subject index

What did I learn About Stata (continued)?

- From https://www.stata-press.com/data/mus2.html it takes three commands to download programs ,datasets and addon ado files.



What did I learn About Stata (continued)?

- Part of mus217binary.do

```
********** 17.4 EXAMPLE (LOGIT, PROBIT, OLS AND GLM MODELS)

* Read in data, define globals, and summarize key variables
qui use mus217hrs
global xlist age hstatusg hhincome educyear married hisp
global extralist female white chronic adl sretire
summarize ins retire $xlist $extralist

* Logit regression
logit ins retire $xlist, vce(robust)

* Comparison of estimates for logit, probit and LPM models
qui logit ins retire $xlist
estimates store blogit
qui probit ins retire $xlist
estimates store bprobit
qui regress ins retire $xlist
estimates store bols
qui logit ins retire $xlist, vce(robust)
estimates store blogitr
qui probit ins retire $xlist, vce(robust)
estimates store bprobitr
qui regress ins retire $xlist, vce(robust)
estimates store bolsr

* Table for comparing models
estimates table blogit blogitr bprobit bprobitr bols bolsr, ///
    t stats(N ll) b(%7.3f) stfmt(%8.2f) eq(1)

* Wald test for no interactions
global intlist c.age#c.age c.age#i.hstatusg c.age#c.hhincome ///
```

What did I learn About Stata (continued)?

- Use scripts (do files) everywhere
  - especially for original dataset creation
  - and put comments in everywhere.
- You can learn a lot about methods by trying them in Stata
  - go to the commands at the end of the help file
    - ★ read in the data and execute the commands
  - read the Methods and Formulas in the pdf manual
  - if a Stata add-on download the associated dataset.

What did I learn About Stata (continued)?

- Following is at bottom after `help Stata`

Examples

Setup

```
. webuse lbw
```

Logistic regression

```
. logit low age lwt i.race smoke ptl ht ui
. logit, level(99)
```

Setup

```
. webuse nhanes2d
. svyset
```

Logistic regression using survey data

```
. svy: logit highbp height weight age female
```

What did I learn About Stata (continued)?

- If you can do it in Stata then do it in Stata
  - it's usually easier and life is short.
- If you can't, consider Python
  - as R has more overlap with Stata
  - and Python can be called from Stata and vice-versa.

5. What did I learn about Microeconometrics?

- It is too broad to cover in one book
  - this I learnt after the fact
  - volume 1 can be used as a text at masters level or advanced undergrad
  - volume 2 has most of the more advanced / current topics.
- Current emphasis of applied microeconometrics research is on quasi-experimental methods
  - for mostly linear models under minimal assumptions
  - and may be necessary as we become more aware of limitations of some of the quasi-experimental methods.
- But there is still a role to have knowledge of nonlinear models and more parametric methods.
- The following discussion sequentially goes through chapters.

- VOLUME 1
  - intended to be useful as a stand-alone econometrics text
- 1. Stata basics
- 2. Data management and graphics
  - 2D graphs if possible are really useful
    - ★ 3D graphs (not provided by Stata) are generally hard to read.
- 3. Linear regression basics
  - before running a regression, look carefully at your data
    - ★ use `summarize` and e.g. `list in 1/10, clean`.
  - it can be insightful to estimate by OLS ahead of other methods.
- 4. Linear regression extensions

- 5. Simulation
  - extraordinarily useful
    - ★ e.g. placebo check using one's own data.
- 6. Linear regression with correlated errors
  - the few clusters problem for cluster-robust inference is more common than originally thought
    - ★ with no universal solution
    - ★ use addon `boottest` for Wild cluster bootstrap
    - ★ most applied statistics instead use mixed linear models
  - survey commands (svy: in Stata) are seldom used in econometrics which is reasonable since
    - ★ weighting - we usually don't weight (but can e.g. `regress y x [aw=pop]`)
    - ★ clustering - we use option `vce(cluster)`
    - ★ stratification - ignore with possible loss of efficiency.

- 7. Linear instrumental variables regression
  - there is a large literature on weak instruments
  - people ignore the result that if the instrument is slightly correlated with the error then IV can be more inconsistent than OLS
  - for just-identified single endogenous case with weak instruments
    - ★ do not use first-stage F statistic as a pretest
    - ★ problems can rise even if F much greater than 10
    - ★ just directly use Anderson-Rubin test (which can robustify).

- 8. Linear panel-data models: Basics
  - straightforward but FE can be much less efficient
  - useful is correlated random effect model or Mundlak approach
    - ★ suppose $\alpha_i = \bar{\mathbf{x}}_i'\gamma + \eta_i$ then $y_{it} = \mathbf{x}_i'\beta + \bar{\mathbf{x}}_i'\gamma + (\eta_i + \varepsilon_{it})$
    - ★ can use for nonlinear panel and two-way fixed effects.
- 9. Linear panel-data models: Extensions
  - Arellano-Bond
  - long panels with $N \longrightarrow \infty$ and $T \longrightarrow \infty$ allow richer models such as interactive effects.

- 10. Introduction to nonlinear regression
  - key models are logit, probit, poisson (exponential mean), NLS.
- 11. Tests of hypotheses and model selection
  - Wald test generally used
  - power of tests is often very low
    - ★ e.g. if $\widehat{\beta} \overset{a}{\sim} N(0,1)$ then a 5% test of $H_0 : \beta = 0$ against $H_0 : \beta = 1.96$ has power 0.50.
  - adjust p-values if testing subgroups or multiple outcomes.
- 12. Bootstrap methods
  - usually used to get standard errors with no asymptotic refinement
  - leading exception is the Wild cluster bootstrap for few clusters
    - ★ use Stata addon `boottest`.

- 13. Nonlinear regression methods.
  - use marginal effects as $d\mathrm{E}[y|x]/dx \neq \beta$ in a nonlinear model
    - ★ distinguish between AME and MEM
    - ★ and between finite difference (use factor variable notation) and calculus MEs
    - ★ `margins` and `margins, dydx(*)` are especially useful
  - endogeneity becomes more difficult in a nonlinear model
    - ★ there is more than one way to bring in endogeneity leading to differing estimates
    - ★ and one cannot use the usual two-stage LS interpretation of 2SLS.

- 14. Flexible regression: finite mixtures and nonparametric
  - two-component finite mixture model can work well
    - ★ density $f(y|\mathbf{x},\boldsymbol{\beta}) = \pi f_1(y|\mathbf{x},\boldsymbol{\beta}_1) + (1-\pi) f_2(y|\mathbf{x},\boldsymbol{\beta}_2)$.
- 15. Quantile regression
  - most people do conditional quantile regression
    - ★ which only considers quantiles of the error term $y - E[y|\mathbf{x}]$
  - but usually we are interested in unconditional quantile regression
    - ★ quantiles of y e.g. effect of change x at various earnings levels.
- Appendices
  - A: Programming in Stata
  - B: Mata
  - C: Optimization in Mata.

- 16. Nonlinear optimization methods
  - key to understand is Newton-Raphson algorithm
  - also stochastic gradient descent (used in machine learning).
- 17. Binary outcome models
  - logit and probit can be used for fractional data $(0 \leq y \leq 1)$
    - ★ but then use robust standard errors.
- 18. Multinomial models
  - multinomial logit is very restrictive
  - most flexible multinomial probit is difficult to estimate
  - random parameters logit is more popular.

- 19. Tobit and selection models
  - enormous challenge as with censoring or truncation the sample is not representative of the population
    - ★ e.g. observe only y given $y > 0$ but want to model $-\infty < y < \infty$
  - so need to be highly parametric
    - ★ there are commands but they rely a lot on assumptions on observables (e.g. i.i.d. normal errors)
  - or restrict analysis to settings with plausible natural experiments.

- 20. Count-data models
  - Poisson regression can be used not just for counts
  - so use whenever happy to specify $E[y|\mathbf{x}] = \exp(\mathbf{x}'\boldsymbol{\beta})$
    - ★ good for data with $y \geq 0$
    - ★ can interpret β as a semi-elasticity
    - ★ but do use robust SE's
  - to model probabilities $\Pr[y = j|\mathbf{x}]$ need to use a richer model
    - ★ such as negative binomial or a hurdle model.
- 21. Survival analysis for duration data
  - straightforward if spells completely observed
  - but in practice they are censored (e.g. some are incomplete)
  - can be parametric e.g. Weibull
  - or use Cox semiparametric proportional hazards model
    - ★ this models the hazard rate Pr[die at time t|not yet dead]
    - ★ unlike the conditional mean this does not require completed spells.

- 22. Nonlinear panel models
  - due to the incidental parameters problem consistent estimation of fixed effects models in short panels is only possible for
    - ★ linear models (more generally model with additive errors)
    - ★ Poisson model (exponential conditional mean)
    - ★ logit model
  - in short panels one can use a bias-adjusted estimator
  - or use correlated random effects (Mundlak approach).
- 23. Parametric models for heterogeneity and endogeneity
  - finite mixture models, linear and nonlinear mixed effect models, generalized structural equation models, ERM commands
  - these are not currently in favor in econometrics.

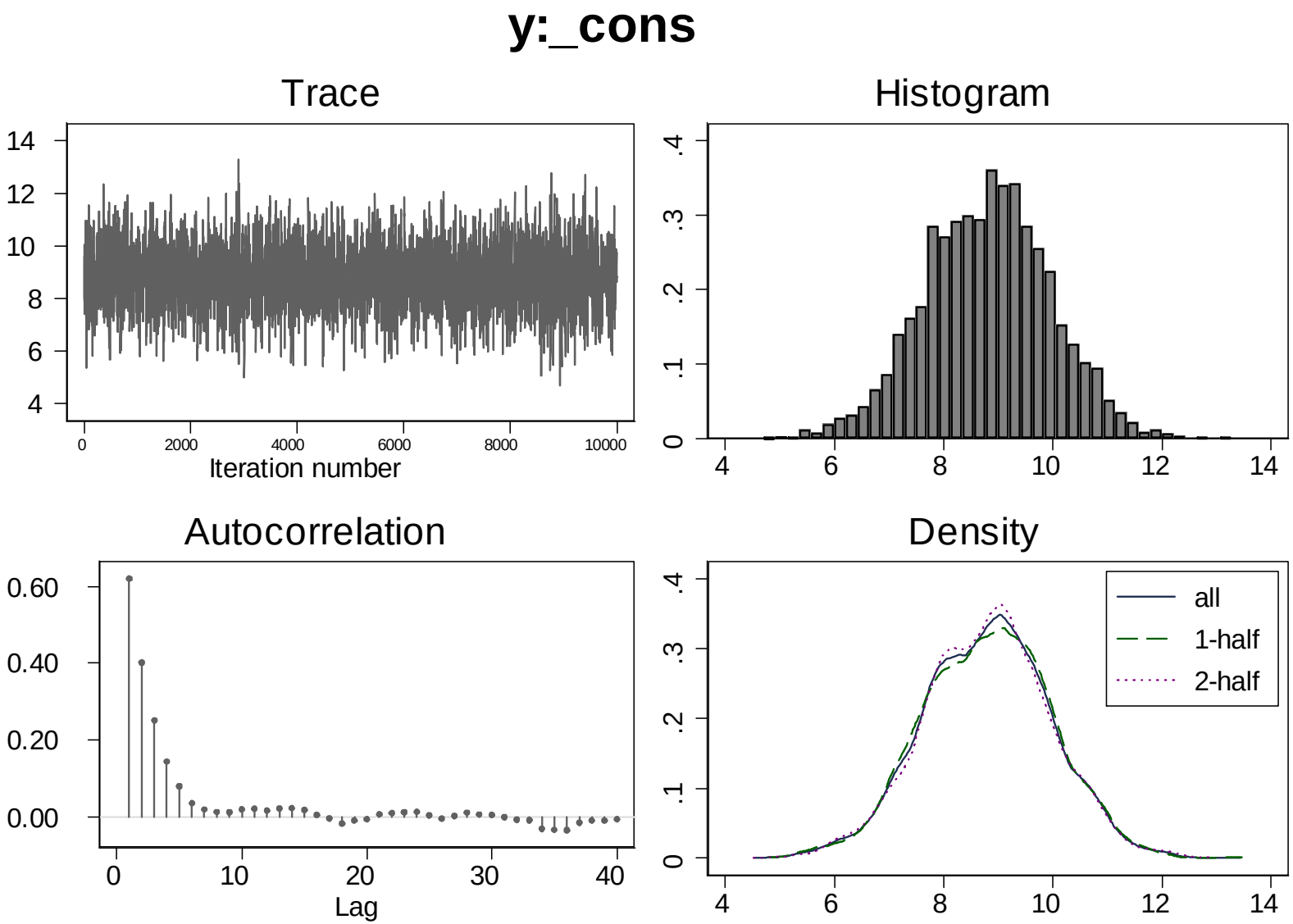
- 24. RCTs and exogenous treatment effects
  - under either random assignment or the crucial assumption of unconfoundedness (selection on observables only)
  - methods are regression adjustment, inverse-probability weighting, doubly robust IPW, matching
  - coverage here is quite complete.
- 25. Endogenous treatment effects
  - parametric approaches use ERM and ET commands and assume common treatment effect parameter
    - ★ which still implies heterogeneous effect if the model is nonlinear.
  - LATE is difficult to extend beyond binary treatment and binary instrument
  - differences-in-differences is active area with staggered treatment
  - synthetic control has challenge in doing inference
  - regression discontinuity design looks solid - use `rdrobust`
  - quantile regression with endogenous treatment is difficult
  - this chapter is introductory and many methods are still being researched.

- 26. Spatial regression
  - spatial in error is no problem - $y_i = \mathbf{x}_i'\boldsymbol{\beta}+$ spatial error
  - spatial in mean requires specifying $\mathbf{W}$ matrix in $\mathbf{y} = \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$
    - ★ e.g. peer effects
    - ★ estimate by IV has less assumptions than MLE.
- 27. Semiparametric regression
  - $E[y|\mathbf{x}] = f(\mathbf{x})$ with $f(\cdot)$ unspecified
  - suffers from curse of dimensionality
  - `npregress` gives average effect & plot for different values of a single x
  - or use semiparametric model that reduces unknown dimension to one
    - ★ partial linear model $\beta x_1 + g(\mathbf{x}_2)$
    - ★ single index model $g(\mathbf{x}'\boldsymbol{\beta})$
    - ★ generalized additive model $g_1(x_1) + g_2(x_2) + \cdots$
  - not popular and now instead use machine learning
    - ★ though curse of dimensionality is still relevant for ML.

- 28. Machine learning for prediction and inference
  - if you only read one chapter this is the one to read
  - machine learning (ML) as the computer learns from data
    - ★ rather than use a model specified by the researcher
  - supervised learning has both y and $\mathbf{x}$
    - ★ regression - y is continuous
    - ★ classification - y is discrete
  - unsupervised learning has only $\mathbf{x}$ e.g. principal components analysis.
  - ML is mostly used for prediction
    - ★ trade-off between bias and variance (so not unbiased)
    - ★ k-fold cross validation use out-of-sample prediction to assess models
    - ★ MLs include lasso, ridge, regression trees, random forests and neural networks.
  - econometricians also interested in causal inference
    - ★ e.g. $E[y|x,\mathbf{z}] = \alpha x + \mathbf{z}'\gamma$ and lasso used to choose $\mathbf{z}$
    - ★ uses a special orthogonal moment condition and cross fitting.

- 29. Bayesian methods: Basics
  - it is important to know Bayesian methods (often not taught)
  - key is combine data with a prior on the parameters
    - ★ $p(\boldsymbol{\theta}|\mathbf{y},\mathbf{X}) \propto L(\mathbf{y}|\boldsymbol{\theta},\mathbf{X}) \times \pi(\boldsymbol{\theta})$
    - ★ Posterior $\propto$ Likelihood $\times$ Prior
  - in some applications we have an informative prior
  - but in regression applications usually an uninformative prior
  - Markov chain Monte Carlo (MCMC) provides a way to get (correlated) draws from the posterior even if the posterior is intractable!
    - ★ especially useful alternative to gradient methods for tough maximum likelihood estimation (and scales well)
  - once we have the draws from the posterior for θ we are done
    - ★ no asymptotic theory.
- 30. Bayesian methods: MCMC algorithms
  - codes from scratch Metropolis-Hastings algorithm for probit
  - codes from scratch data augmentation and Gibbs sampler for probit
  - multiple imputation is not used much in econometrics.

- Bayesian example summarizing the (correlated) MCMC draws of β



What additional topics should have been included?

- More references
  - for space reasons only about 240 are given.
- Mediation analysis
  - an oversight - it is not a difficult topic to explain.
- Bounds under partial identification
  - this would have taken more time.
- Robustness checks for a research paper
  - or at least reference to good example papers.
- More on endogenous treatment
  - but this really is a separate book and currently is a moving target.

6. Will I make money?

- Not directly as book sales are now very low (in part due to piracy).
- Potentially indirectly through publication record / visibility.
- And people do thank me for books such as this.