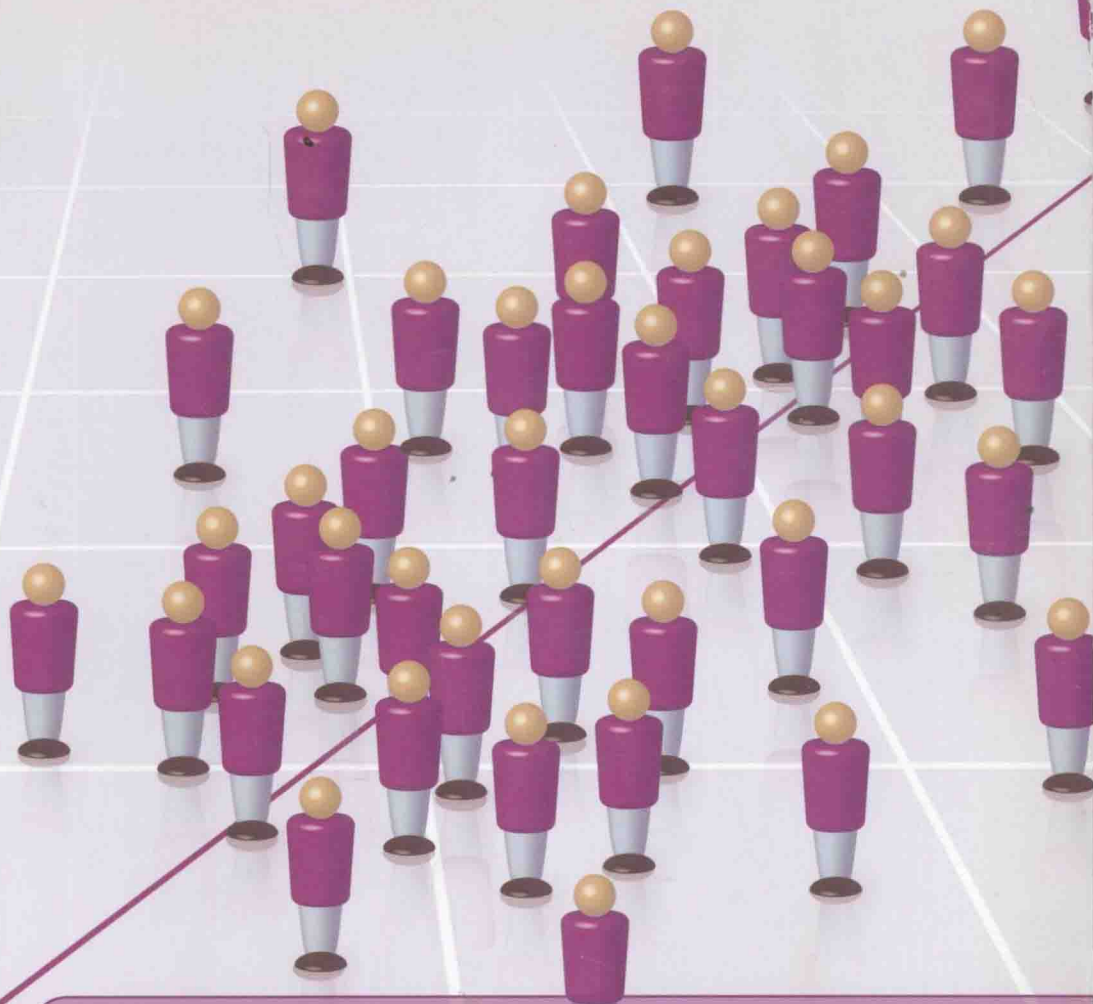




Microeconometrics Using Stata

Revised Edition

**A. COLIN CAMERON
PRAVIN K. TRIVEDI**

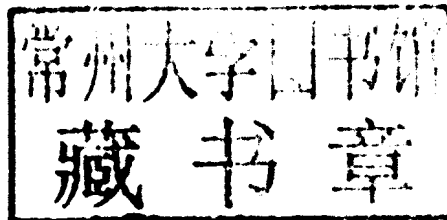
Stata
Press

Microeconometrics Using Stata

Revised Edition

A. COLIN CAMERON
Department of Economics
University of California
Davis, CA

PRAVIN K. TRIVEDI
Department of Economics
Indiana University
Bloomington, IN



A Stata Press Publication
StataCorp LP
College Station, Texas



Copyright © 2009, 2010 by StataCorp LP
All rights reserved. First edition 2009
Revised edition 2010

Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845

Typeset in L^AT_EX 2_ε

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

ISBN-10: 1-59718-073-4

ISBN-13: 978-1-59718-073-3

No part of this book may be reproduced, stored in a retrieval system, or transcribed, in any form or by any means—electronic, mechanical, photocopy, recording, or otherwise—without the prior written permission of StataCorp LP.

Stata is a registered trademark of StataCorp LP. L^AT_EX 2_ε is a trademark of the American Mathematical Society.

Microeconometrics Using Stata

Revised Edition

Preface to the Revised Edition

Microeconometrics Using Stata, published in December 2008, was written for Stata 10.1. The book incorporated version 10.1 additions to Stata 10.0, most notably, the new random-number generators.

In this revised edition, we present other additions to Stata 10 that appear for the first time in Stata 11. With few exceptions, we present these additions in a way that reproduces the results given in the first edition.

First, we introduce the new construct of factor variables. These provide a simple way to specify models with sets of indicator variables formed from a categorical variable and to specify models with interactions. Factor variables replace the `xi` prefix command. See especially section 1.3.4 and the end of section 2.4.7.

Second, we describe the new `margins` command for prediction and for computation of marginal effects in regression models. The `margins` command with options including the `dydx()` option replaces the Stata `mf` command and the user-written `margeff` command. Additionally, the `margins` command when used in conjunction with factor variables can simplify computation of marginal effects in models with interactions. See sections 10.5 and 10.6, especially subsections 10.5.7 and 10.6.5. Throughout this revised edition, notably, in chapters 14–17, we replace `mf` and `margeff` with the `margins` command.

In the first edition, we most often calculated the marginal effect at the mean (MEM), rather than the average marginal effect (AME), because the `mf` command did not compute the AME. The new `margins` command can compute both the MEM and the AME. In this revised edition, we have endeavored to replicate the results given in the first edition. For that reason, we continue to most frequently calculate the MEM, though in practice, the AME is usually preferred.

Third, we describe the new `gmm` command for generalized method of moments and nonlinear instrumental-variables estimation. See sections 10.3.8 and 17.5.2.

Fourth, we present some minor changes that need to be made to the existing `ml` command when the `d1` and `d2` methods are used. These changes arise because the `ml` command is now a front-end to the new Mata `moptimize()` function. We also present the new `lf0`, `lf1`, and `lf2` methods. See section 11.6. The Mata `optimize()` evaluator has been renamed to `gf` evaluator; see section 11.7.

We thank the Stata staff, especially Patricia Branton, David Drukker, Lisa Gilmore, Deirdre Patterson, and Brian Poi, for their assistance in preparing this revised edition.

Davis, CA
Bloomington, IN
January 2010

A. Colin Cameron
Pravin K. Trivedi

Preface to the First Edition

This book explains how an econometrics computer package, Stata, can be used to perform regression analysis of cross-section and panel data. The term microeconomics is used in the book title because the applications are to economics-related data and because the coverage includes methods such as instrumental-variables regression that are emphasized more in economics than in some other areas of applied statistics. However, many issues, models, and methodologies discussed in this book are also relevant to other social sciences.

The main audience is graduate students and researchers. For them, this book can be used as an adjunct to our own *Microeconomics: Methods and Applications* (Cameron and Trivedi 2005), as well as to other graduate-level texts such as Greene (2008) and Wooldridge (2002). By comparison to these books, we present little theory and instead emphasize practical aspects of implementation using Stata. More advanced topics we cover include quantile regression, weak instruments, nonlinear optimization, bootstrap methods, nonlinear panel-data methods, and Stata's matrix programming language, Mata.

At the same time, the book provides introductions to topics such as ordinary least-squares regression, instrumental-variables estimation, and logit and probit models so that it is suitable for use in an undergraduate econometrics class, as a complement to an appropriate undergraduate-level text. The following table suggests sections of the book for an introductory class, with the caveat that in places formulas are provided using matrix algebra.

Stata basics	Chapter 1.1–1.4
Data management	Chapter 2.1–2.4, 2.6
OLS	Chapter 3.1–3.6
Simulation	Chapter 4.6–4.7
GLS (heteroskedasticity)	Chapter 5.3
Instrumental variables	Chapter 6.2–6.3
Linear panel data	Chapter 8
Logit and probit models	Chapter 14.1–14.4
Tobit model	Chapter 16.1–16.3

Although we provide considerable detail on Stata, the treatment is by no means complete. In particular, we introduce various Stata commands but avoid detailed listing and description of commands as they are already well documented in the Stata manuals

and online help. Typically, we provide a pointer and a brief discussion and often an example.

As much as possible, we provide template code that can be adapted to other problems. Keep in mind that to shorten output for this book, our examples use many fewer regressors than necessary for serious research. Our code often suppresses intermediate output that is important in actual research, because of extensive use of command `quietly` and options `nolog`, `nodots`, and `noheader`. And we minimize the use of graphs compared with typical use in exploratory data analysis.

We have used Stata 10, including Stata updates.¹ Instructions on how to obtain the datasets and the do-files used in this book are available on the Stata Press web site at <http://www.stata-press.com/data/mus.html>. Any corrections to the book will be documented at <http://www.stata-press.com/books/mus.html>.

We have learned a lot of econometrics, in addition to learning Stata, during this project. Indeed, we feel strongly that an effective learning tool for econometrics is hands-on learning by opening a Stata dataset and seeing the effect of using different methods and variations on the methods, such as using robust standard errors rather than default standard errors. This method is beneficial at all levels of ability in econometrics. Indeed, an efficient way of familiarizing yourself with Stata's leading features might be to execute the commands in a relevant chapter on your own dataset.

We thank the many people who have assisted us in preparing this book. The project grew out of our 2005 book, and we thank Scott Parris for his expert handling of that book. Juan Du, Qian Li, and Abhijit Ramalingam carefully read many of the book chapters. Discussions with John Daniels, Oscar Jorda, Guido Kuersteiner, and Doug Miller were particularly helpful. We thank Deirdre Patterson for her excellent editing and Lisa Gilmore for managing the L^AT_EX formatting and production of this book. Most especially, we thank David Drukker for his extensive input and encouragement at all stages of this project, including a thorough reading and critique of the final draft, which led to many improvements in both the econometrics and Stata components of this book. Finally, we thank our respective families for making the inevitable sacrifices as we worked to bring this multiyear project to completion.

Davis, CA
Bloomington, IN
October 2008

A. Colin Cameron
Pravin K. Trivedi

1. To see whether you have the latest update, type `update query`. For those with earlier versions of Stata, some key changes are the following: Stata 9 introduced the matrix programming language, Mata. The syntax for Stata 10 uses the `vce(robust)` option rather than the `robust` option to obtain robust standard errors. A mid-2008 update of version 10 introduced new random-number functions, such as `runiform()` and `rnormal()`.

Contents

	List of tables	xxxv
	List of figures	xxxvii
	Preface to the Revised Edition	xxxix
	Preface to the First Edition	xli
1	Stata basics	1
1.1	Interactive use	1
1.2	Documentation	2
1.2.1	Stata manuals	2
1.2.2	Additional Stata resources	3
1.2.3	The help command	3
1.2.4	The search, findit, and hsearch commands	4
1.3	Command syntax and operators	5
1.3.1	Basic command syntax	5
1.3.2	Example: The summarize command	6
1.3.3	Example: The regress command	7
1.3.4	Factor variables	9
1.3.5	Abbreviations, case sensitivity, and wildcards	11
1.3.6	Arithmetic, relational, and logical operators	12
1.3.7	Error messages	12
1.4	Do-files and log files	13
1.4.1	Writing a do-file	13
1.4.2	Running do-files	14
1.4.3	Log files	14
1.4.4	A three-step process	15

1.4.5	Comments and long lines	16
1.4.6	Different implementations of Stata	17
1.5	Scalars and matrices	17
1.5.1	Scalars	17
1.5.2	Matrices	18
1.6	Using results from Stata commands	18
1.6.1	Using results from the r-class command summarize	18
1.6.2	Using results from the e-class command regress	19
1.7	Global and local macros	21
1.7.1	Global macros	21
1.7.2	Local macros	22
1.7.3	Scalar or macro?	23
1.8	Looping commands	24
1.8.1	The foreach loop	25
1.8.2	The forvalues loop	26
1.8.3	The while loop	26
1.8.4	The continue command	27
1.9	Some useful commands	27
1.10	Template do-file	27
1.11	User-written commands	28
1.12	Stata resources	29
1.13	Exercises	29
2	Data management and graphics	31
2.1	Introduction	31
2.2	Types of data	31
2.2.1	Text or ASCII data	32
2.2.2	Internal numeric data	32
2.2.3	String data	33
2.2.4	Formats for displaying numeric data	33

2.3	Inputting data	34
2.3.1	General principles	34
2.3.2	Inputting data already in Stata format	35
2.3.3	Inputting data from the keyboard	36
2.3.4	Inputting nontext data	36
2.3.5	Inputting text data from a spreadsheet	37
2.3.6	Inputting text data in free format	38
2.3.7	Inputting text data in fixed format	38
2.3.8	Dictionary files	39
2.3.9	Common pitfalls	39
2.4	Data management	40
2.4.1	PSID example	40
2.4.2	Naming and labeling variables	43
2.4.3	Viewing data	44
2.4.4	Using original documentation	45
2.4.5	Missing values	45
2.4.6	Imputing missing data	47
2.4.7	Transforming data (generate, replace, egen, recode)	48
	The generate and replace commands	48
	The egen command	49
	The recode command	49
	The by prefix	49
	Indicator variables	50
	Set of indicator variables	50
	Interactions	51
	Demeaning	52
2.4.8	Saving data	52
2.4.9	Selecting the sample	53
2.5	Manipulating datasets	54
2.5.1	Ordering observations and variables	55

2.5.2	Preserving and restoring a dataset	55
2.5.3	Wide and long forms for a dataset	55
2.5.4	Merging datasets	56
2.5.5	Appending datasets	58
2.6	Graphical display of data	58
2.6.1	Stata graph commands	59
	Example graph commands	59
	Saving and exporting graphs	60
	Learning how to use graph commands	61
2.6.2	Box-and-whisker plot	61
2.6.3	Histogram	63
2.6.4	Kernel density plot	63
2.6.5	Twoway scatterplots and fitted lines	66
2.6.6	Lowess, kernel, local linear, and nearest-neighbor regression	67
2.6.7	Multiple scatterplots	69
2.7	Stata resources	70
2.8	Exercises	70
3	Linear regression basics	73
3.1	Introduction	73
3.2	Data and data summary	73
3.2.1	Data description	73
3.2.2	Variable description	74
3.2.3	Summary statistics	75
3.2.4	More-detailed summary statistics	76
3.2.5	Tables for data	77
3.2.6	Statistical tests	80
3.2.7	Data plots	80
3.3	Regression in levels and logs	81
3.3.1	Basic regression theory	81
3.3.2	OLS regression and matrix algebra	82

3.3.3	Properties of the OLS estimator	83
3.3.4	Heteroskedasticity-robust standard errors	84
3.3.5	Cluster-robust standard errors	84
3.3.6	Regression in logs	85
3.4	Basic regression analysis	86
3.4.1	Correlations	86
3.4.2	The regress command	87
3.4.3	Hypothesis tests	88
3.4.4	Tables of output from several regressions	89
3.4.5	Even better tables of regression output	90
3.4.6	Factor variables for categorical variables and interactions . .	92
3.5	Specification analysis	94
3.5.1	Specification tests and model diagnostics	94
3.5.2	Residual diagnostic plots	95
3.5.3	Influential observations	96
3.5.4	Specification tests	97
	Test of omitted variables	98
	Test of the Box-Cox model	98
	Test of the functional form of the conditional mean	99
	Heteroskedasticity test	100
	Omnibus test	102
3.5.5	Tests have power in more than one direction	102
3.6	Prediction	104
3.6.1	In-sample prediction	104
3.6.2	MEs and elasticities	106
3.6.3	Prediction in logs: The retransformation problem	108
3.6.4	Prediction exercise	109
3.7	Sampling weights	111
3.7.1	Weights	111
3.7.2	Weighted mean	112

3.7.3	Weighted regression	113
3.7.4	Weighted prediction and MEs	114
3.8	OLS using Mata	115
3.9	Stata resources	117
3.10	Exercises	117
4	Simulation	119
4.1	Introduction	119
4.2	Pseudorandom-number generators: Introduction	120
4.2.1	Uniform random-number generation	120
4.2.2	Draws from normal	122
4.2.3	Draws from t, chi-squared, F, gamma, and beta	123
4.2.4	Draws from binomial, Poisson, and negative binomial	124
	Independent (but not identically distributed) draws from binomial	124
	Independent (but not identically distributed) draws from Poisson	125
	Histograms and density plots	126
4.3	Distribution of the sample mean	127
4.3.1	Stata program	128
4.3.2	The simulate command	129
4.3.3	Central limit theorem simulation	129
4.3.4	The postfile command	130
4.3.5	Alternative central limit theorem simulation	131
4.4	Pseudorandom-number generators: Further details	131
4.4.1	Inverse-probability transformation	132
4.4.2	Direct transformation	133
4.4.3	Other methods	133
4.4.4	Draws from truncated normal	134
4.4.5	Draws from multivariate normal	135
	Direct draws from multivariate normal	135
	Transformation using Cholesky decomposition	136

4.4.6	Draws using Markov chain Monte Carlo method	136
4.5	Computing integrals	138
4.5.1	Quadrature	139
4.5.2	Monte Carlo integration	139
4.5.3	Monte Carlo integration using different S	140
4.6	Simulation for regression: Introduction	141
4.6.1	Simulation example: OLS with χ^2 errors	141
4.6.2	Interpreting simulation output	144
	Unbiasedness of estimator	144
	Standard errors	144
	t statistic	144
	Test size	145
	Number of simulations	146
4.6.3	Variations	146
	Different sample size and number of simulations	146
	Test power	146
	Different error distributions	147
4.6.4	Estimator inconsistency	147
4.6.5	Simulation with endogenous regressors	148
4.7	Stata resources	150
4.8	Exercises	150
5	GLS regression	153
5.1	Introduction	153
5.2	GLS and FGLS regression	153
5.2.1	GLS for heteroskedastic errors	153
5.2.2	GLS and FGLS	154
5.2.3	Weighted least squares and robust standard errors	155
5.2.4	Leading examples	155
5.3	Modeling heteroskedastic data	156
5.3.1	Simulated dataset	156

5.3.2	OLS estimation	157
5.3.3	Detecting heteroskedasticity	158
5.3.4	FGLS estimation	160
5.3.5	WLS estimation	162
5.4	System of linear regressions	162
5.4.1	SUR model	162
5.4.2	The sureg command	163
5.4.3	Application to two categories of expenditures	164
5.4.4	Robust standard errors	166
5.4.5	Testing cross-equation constraints	167
5.4.6	Imposing cross-equation constraints	168
5.5	Survey data: Weighting, clustering, and stratification	169
5.5.1	Survey design	170
5.5.2	Survey mean estimation	173
5.5.3	Survey linear regression	173
5.6	Stata resources	175
5.7	Exercises	175
6	Linear instrumental-variables regression	177
6.1	Introduction	177
6.2	IV estimation	177
6.2.1	Basic IV theory	177
6.2.2	Model setup	179
6.2.3	IV estimators: IV, 2SLS, and GMM	180
6.2.4	Instrument validity and relevance	181
6.2.5	Robust standard-error estimates	182
6.3	IV example	183
6.3.1	The ivregress command	183
6.3.2	Medical expenditures with one endogenous regressor	184
6.3.3	Available instruments	185
6.3.4	IV estimation of an exactly identified model	186

6.3.5	IV estimation of an overidentified model	187
6.3.6	Testing for regressor endogeneity	188
6.3.7	Tests of overidentifying restrictions	191
6.3.8	IV estimation with a binary endogenous regressor	192
6.4	Weak instruments	194
6.4.1	Finite-sample properties of IV estimators	194
6.4.2	Weak instruments	195
	Diagnostics for weak instruments	195
	Formal tests for weak instruments	196
6.4.3	The estat firststage command	197
6.4.4	Just-identified model	197
6.4.5	Overidentified model	199
6.4.6	More than one endogenous regressor	200
6.4.7	Sensitivity to choice of instruments	200
6.5	Better inference with weak instruments	202
6.5.1	Conditional tests and confidence intervals	202
6.5.2	LIML estimator	204
6.5.3	Jackknife IV estimator	204
6.5.4	Comparison of 2SLS, LIML, JIVE, and GMM	205
6.6	3SLS systems estimation	206
6.7	Stata resources	208
6.8	Exercises	208
7	Quantile regression	211
7.1	Introduction	211
7.2	QR	211
7.2.1	Conditional quantiles	212
7.2.2	Computation of QR estimates and standard errors	213
7.2.3	The qreg, bsqreg, and sqreg commands	213
7.3	QR for medical expenditures data	214
7.3.1	Data summary	214