

Microeconometrics Using Stata

Second Edition

A. COLIN CAMERON

Department of Economics

University of California, Davis, CA

and

School of Economics

University of Sydney, Sydney, Australia

PRAVIN K. TRIVEDI

School of Economics

University of Queensland, Brisbane, Australia

and

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.

Contents

	List of tables	xvii
	List of figures	xix
	Preface to the Second Edition	xxiii
	Preface to the First Edition	xxv
1	Stata basics	1
	1.1 Interactive use	1
	1.2 Documentation	2
	1.3 Command syntax and operators	5
	1.4 Do-files and log files	13
	1.5 Scalars and matrices	17
	1.6 Using results from Stata commands	18
	1.7 Global and local macros	21
	1.8 Looping commands	24
	1.9 Mata and Python in Stata	27
	1.10 Some useful commands	27
	1.11 Template do-file	27
	1.12 Community-contributed commands	28
	1.13 Additional resources	29
	1.14 Exercises	29
2	Data management and graphics	31
	2.1 Introduction	31
	2.2 Types of data	31
	2.3 Inputting data	34
	2.4 Data management	41

2.5	Manipulating datasets	57
2.6	Graphical display of data	64
2.7	Additional resources	77
2.8	Exercises	78
3	Linear regression basics	81
3.1	Introduction	81
3.2	Data and data summary	81
3.3	Transformation of data before regression	89
3.4	Linear regression	91
3.5	Basic regression analysis	97
3.6	Specification analysis	114
3.7	Specification tests	124
3.8	Sampling weights	131
3.9	OLS using Mata	135
3.10	Additional resources	137
3.11	Exercises	137
4	Linear regression extensions	139
4.1	Introduction	139
4.2	In-sample prediction	139
4.3	Out-of-sample prediction	147
4.4	Predictive margins	150
4.5	Marginal effects	163
4.6	Regression decomposition analysis	173
4.7	Shapley decomposition of relative regressor importance	179
4.8	Differences-in-differences estimators	181
4.9	Additional resources	188
4.10	Exercises	188
5	Simulation	191
5.1	Introduction	191
5.2	Pseudorandom-number generators	192

5.3	Distribution of the sample mean	198
5.4	Pseudorandom-number generators: Further details	203
5.5	Computing integrals	210
5.6	Simulation for regression: Introduction	215
5.7	Additional resources	225
5.8	Exercises	225
6	Linear regression with correlated errors	227
6.1	Introduction	227
6.2	GLS and FGLS regression	228
6.3	Modeling heteroskedastic data	232
6.4	OLS for clustered data	238
6.5	FGLS estimators for clustered data	246
6.6	Fixed effects estimator for clustered data	250
6.7	Linear mixed models for clustered data	256
6.8	Systems of linear regressions	265
6.9	Survey data: weighting, clustering, and stratification	273
6.10	Additional resources	279
6.11	Exercises	280
7	Linear instrumental variables regression	283
7.1	Introduction	283
7.2	Simultaneous equations model	284
7.3	IV estimation	288
7.4	IV example	294
7.5	Weak instruments	307
7.6	Diagnostics and tests for weak instruments	316
7.7	Inference with weak instruments	329
7.8	Finite sample inference with weak instruments	337
7.9	Other estimators	338
7.10	3SLS systems estimation	341
7.11	Additional resources	343

7.12	Exercises	343
8	Linear panel-data models: basics	347
8.1	Introduction	347
8.2	Panel-data methods overview	347
8.3	Summary of panel-data	353
8.4	Pooled or population-averaged estimators	366
8.5	FE or within estimator	369
8.6	Between estimator	374
8.7	RE estimator	375
8.8	Comparison of estimators	378
8.9	First-difference estimator	384
8.10	Panel-data management	386
8.11	Additional resources	390
8.12	Exercises	390
9	Linear panel-data models: extensions	393
9.1	Introduction	393
9.2	Panel IV estimation	393
9.3	Hausman–Taylor estimator	396
9.4	Arellano–Bond estimator	399
9.5	Long panels	414
9.6	Additional resources	424
9.7	Exercises	424
10	Introduction to nonlinear regression	427
10.1	Introduction	427
10.2	Binary outcome models	427
10.3	Probit model	430
10.4	Marginal effects and coefficient interpretation	433
10.5	Logit model	439
10.6	Nonlinear least squares	440
10.7	Other nonlinear estimators	442

10.8	Additional resources	443
10.9	Exercises	443
11	Tests of hypotheses and model specification	445
11.1	Introduction	445
11.2	Critical values and p-values	446
11.3	Wald tests and confidence intervals	450
11.4	Likelihood-ratio tests	463
11.5	Lagrange multiplier test (or score test)	467
11.6	Multiple Testing	470
11.7	Test size and power	477
11.8	The power commands for multiple regression	483
11.9	Specification tests	493
11.10	Permutation tests and randomization tests	495
11.11	Additional resources	497
11.12	Exercises	498
12	Bootstrap methods	501
12.1	Introduction	501
12.2	Bootstrap methods	501
12.3	Bootstrap pairs using the vce(bootstrap) option	503
12.4	Bootstrap pairs using the bootstrap command	510
12.5	Percentile-t bootstraps with asymptotic refinement	518
12.6	Wild bootstrap with asymptotic refinement	522
12.7	Bootstrap pairs using bsample and simulate	531
12.8	Alternative resampling schemes	532
12.9	The jackknife	537
12.10	Additional resources	538
12.11	Exercises	539
13	Nonlinear regression methods	541
13.1	Introduction	541
13.2	Nonlinear example: doctor visits	541

13.3	Nonlinear regression methods	544
13.4	Different estimates of the VCE	557
13.5	Prediction	564
13.6	Predictive margins	569
13.7	Marginal effects	572
13.8	Model diagnostics	587
13.9	Clustered data	591
13.10	Additional resources	598
13.11	Exercises	598
14	Flexible regression: finite mixtures and nonparametric	601
14.1	Introduction	601
14.2	Models based on finite mixtures	602
14.3	FMM example: Earnings of doctors	608
14.4	Global polynomials	620
14.5	Regression splines	623
14.6	Nonparametric regression	629
14.7	Partially parametric regression	634
14.8	Additional resources	635
14.9	Exercises	635
15	Quantile regression	637
15.1	Introduction	637
15.2	Conditional quantile regression	638
15.3	Conditional QR for medical expenditures data	641
15.4	Conditional QR for generated heteroskedastic data	653
15.5	Quantile treatment effects for a binary treatment	656
15.6	Additional resources	659
15.7	Exercises	659
16	Nonlinear optimization methods	663
16.1	Introduction	663
16.2	Newton–Raphson method	663

16.3	Gradient methods	668
16.4	Overview of ml, moptimize and optimize commands	672
16.5	The ml command: lf method	674
16.6	Checking the program	680
16.7	The ml command: lf0-lf2, d0-d2 and gf0 methods	686
16.8	Nonlinear IV (GMM) example	693
16.9	Additional resources	696
16.10	Exercises	696
17	Binary outcome models	699
17.1	Introduction	699
17.2	Some parametric models	699
17.3	Estimation	702
17.4	Example	704
17.5	Goodness of fit and prediction	710
17.6	Marginal effects	717
17.7	Clustered data	720
17.8	Additional models	721
17.9	Endogenous regressors	726
17.10	Grouped and aggregate data	734
17.11	Additional resources	737
17.12	Exercises	737
18	Multinomial models	739
18.1	Introduction	739
18.2	Multinomial models overview	739
18.3	Multinomial example: choice of fishing mode	743
18.4	Multinomial logit model	746
18.5	Alternative-specific conditional logit model	751
18.6	Nested logit model	759
18.7	Multinomial probit model	765
18.8	Alternative-specific random-parameters logit	770

18.9	Ordered outcome models	773
18.10	Clustered data	777
18.11	Multivariate outcomes	778
18.12	Additional resources	781
18.13	Exercises	782
19	Tobit and selection models	783
19.1	Introduction	783
19.2	Tobit model	784
19.3	Tobit model example	786
19.4	Tobit for lognormal data	795
19.5	Two-part model in logs	803
19.6	Selection models	806
19.7	Non-normal models of selection	813
19.8	Prediction from models with outcome in logs	817
19.9	Endogenous regressors	820
19.10	Missing data	821
19.11	Panel attrition	826
19.12	Additional resources	847
19.13	Exercises	847
20	Count-data models	849
20.1	Introduction	849
20.2	Modeling strategies for count data	850
20.3	Poisson and negative binomial models	854
20.4	Hurdle model	869
20.5	Finite-mixture models	875
20.6	Zero-inflated models	893
20.7	Endogenous regressors	901
20.8	Clustered data	910
20.9	QR for count data	912
20.10	Additional resources	917

20.11	Exercises	918
21	Survival analysis for duration data	921
21.1	Introduction	921
21.2	Data and data summary	922
21.3	Survivor and hazard functions	926
21.4	Semiparametric regression model	931
21.5	Fully parametric regression models	939
21.6	Multiple-records data	949
21.7	Discrete-time hazards logit model	951
21.8	Time-varying regressors	955
21.9	Clustered data	955
21.10	Additional resources	956
21.11	Exercises	956
22	Nonlinear panel models	959
22.1	Introduction	959
22.2	Nonlinear panel-data overview	959
22.3	Nonlinear panel-data example	964
22.4	Binary outcome and ordered outcome models	967
22.5	Tobit and interval-data models	984
22.6	Count-data models	988
22.7	Panel quantile regression	999
22.8	Endogenous regressors in nonlinear panel models	1001
22.9	Additional resources	1002
22.10	Exercises	1002
23	Parametric models for heterogeneity and endogeneity	1005
23.1	Introduction	1005
23.2	Finite mixtures and unobserved heterogeneity	1006
23.3	Empirical examples of finite mixture models	1008
23.4	Nonlinear mixed effects models	1034
23.5	SEM for linear structural equation models	1041

23.6	Generalized SEM	1060
23.7	ERM commands for endogeneity and selection	1070
23.8	Additional resources	1074
23.9	Exercises	1075
24	RCTs and exogenous treatment effects	1077
24.1	Introduction	1077
24.2	Potential outcomes	1079
24.3	Randomized controlled trials	1080
24.4	Regression in an RCT	1089
24.5	Treatment evaluation with exogenous treatment	1097
24.6	Treatment evaluation methods and estimators	1099
24.7	Stata commands for treatment evaluation	1109
24.8	Oregon Health Insurance Experiment	1111
24.9	Treatment effect estimates using the OHIE data	1118
24.10	Multilevel treatment effects	1127
24.11	Conditional quantile treatment effects	1135
24.12	Additional resources	1137
24.13	Exercises	1138
25	Endogenous treatment effects	1141
25.1	Introduction	1141
25.2	Parametric methods for endogenous treatment	1142
25.3	ERM commands for endogenous treatment	1145
25.4	ET commands for binary endogenous treatment	1152
25.5	The LATE estimator for heterogeneous effects	1160
25.6	Differences-in-differences and synthetic control	1166
25.7	Regression discontinuity design	1170
25.8	Conditional QR with endogenous regressors	1188
25.9	Unconditional quantiles	1194
25.10	Additional resources	1200
25.11	Exercises	1201

26	Spatial regression	1203
26.1	Introduction	1203
26.2	Overview of spatial regression models	1203
26.3	Geospatial data	1205
26.4	The spatial weighting matrix	1208
26.5	OLS regression and test for spatial correlation	1211
26.6	Spatial dependence in the error	1212
26.7	Spatial autoregressive (SAR) models	1214
26.8	Spatial instrumental variables	1224
26.9	Spatial panel-data models	1225
26.10	Additional resources	1226
26.11	Exercises	1226
27	Semiparametric regression	1229
27.1	Introduction	1229
27.2	Kernel regression	1230
27.3	Series regression	1234
27.4	Nonparametric single regressor example	1235
27.5	Nonparametric multiple regressor example	1245
27.6	Partial linear model	1247
27.7	Single-index model	1251
27.8	Generalized additive model	1253
27.9	Additional resources	1255
27.10	Exercises	1256
28	Machine learning for prediction and inference	1259
28.1	Introduction	1259
28.2	Measuring the predictive ability of a model	1260
28.3	Shrinkage Estimators	1270
28.4	Prediction using LASSO, ridge and elasticnet	1275
28.5	Dimension reduction	1285
28.6	Machine learning methods for prediction	1288

28.7	Prediction application	1293
28.8	Machine learning for inference in partial linear model	1297
28.9	Machine learning for inference in other models	1305
28.10	Additional resources	1311
28.11	Exercises	1312
29	Bayesian methods: basics	1315
29.1	Introduction	1315
29.2	Bayesian introductory example	1316
29.3	Bayesian methods overview	1319
29.4	An i.i.d. example	1325
29.5	Linear regression	1335
29.6	A linear regression example	1338
29.7	Modifying the MH algorithm	1345
29.8	Random effects model	1348
29.9	Bayesian model selection	1351
29.10	Bayesian prediction	1353
29.11	Probit example	1356
29.12	Additional resources	1360
29.13	Exercises	1360
30	Bayesian methods: MCMC Algorithms	1363
30.1	Introduction	1363
30.2	User-provided log-likelihood	1363
30.3	Metropolis-Hastings algorithm in Mata	1367
30.4	Data augmentation and the Gibbs sampler in Mata	1373
30.5	Multiple imputation	1378
30.6	Multiple imputation example	1381
30.7	Regression with complete and incomplete data	1382
30.8	Additional resources	1390
30.9	Exercises	1390
A	Programming in Stata	1399

A.1	Stata matrix commands	1399
A.2	Programs	1405
A.3	Program debugging	1411
A.4	Additional resources	1414
B	Mata	1415
B.1	How to run Mata	1415
B.2	Mata matrix commands	1417
B.3	Programming in Mata	1426
B.4	Additional resources	1428
C	Optimization in Mata	1429
C.1	Mata moptimize function	1429
C.2	Mata optimize() function	1438
C.3	Additional resources	1442
	References	1443
	Author index	1467
	Subject index	1473