

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

Contents

	List of tables	v
	List of figures	vii
	Preface to the Second Edition	ix
	Preface to the First Edition	xi
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	189
4.10	Exercises	189
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	245
6.6	Fixed effects estimator for clustered data	249
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	279
7	Linear instrumental variables regression	281
7.1	Introduction	281
7.2	Simultaneous equations model	282
7.3	IV estimation	286
7.4	IV example	292
7.5	Weak instruments	305
7.6	Diagnostics and tests for weak instruments	314
7.7	Inference with weak instruments	327
7.8	Finite sample inference with weak instruments	335
7.9	Other estimators	336
7.10	3SLS systems estimation	339
7.11	Additional resources	341

7.12	Exercises	341
8	Linear panel-data models: basics	345
8.1	Introduction	345
8.2	Panel-data methods overview	345
8.3	Summary of panel-data	351
8.4	Pooled or population-averaged estimators	364
8.5	FE or within estimator	367
8.6	Between estimator	372
8.7	RE estimator	373
8.8	Comparison of estimators	376
8.9	First-difference estimator	382
8.10	Panel-data management	384
8.11	Additional resources	388
8.12	Exercises	388
9	Linear panel-data models: extensions	391
9.1	Introduction	391
9.2	Panel IV estimation	391
9.3	Hausman–Taylor estimator	394
9.4	Arellano–Bond estimator	397
9.5	Long panels	412
9.6	Additional resources	422
9.7	Exercises	422
10	Introduction to nonlinear regression	425
10.1	Introduction	425
10.2	Binary outcome models	425
10.3	Probit model	428
10.4	Marginal effects and coefficient interpretation	431
10.5	Logit model	437
10.6	Nonlinear least squares	438
10.7	Other nonlinear estimators	440

10.8	Additional resources	441
10.9	Exercises	441
11	Tests of hypotheses and model specification	443
11.1	Introduction	443
11.2	Critical values and p-values	444
11.3	Wald tests and confidence intervals	448
11.4	Likelihood-ratio tests	461
11.5	Lagrange multiplier test (or score test)	465
11.6	Multiple Testing	468
11.7	Test size and power	475
11.8	The power onemean command for multiple regression	481
11.9	Specification tests	491
11.10	Permutation tests and randomization tests	493
11.11	Additional resources	495
11.12	Exercises	496
12	Bootstrap methods	499
12.1	Introduction	499
12.2	Bootstrap methods	499
12.3	Bootstrap pairs using the vce(bootstrap) option	501
12.4	Bootstrap pairs using the bootstrap command	508
12.5	Percentile-t bootstraps with asymptotic refinement	516
12.6	Wild bootstrap with asymptotic refinement	520
12.7	Bootstrap pairs using bsample and simulate	529
12.8	Alternative resampling schemes	530
12.9	The jackknife	535
12.10	Additional resources	536
12.11	Exercises	536
13	Nonlinear regression methods	539
13.1	Introduction	539
13.2	Nonlinear example: doctor visits	539

13.3	Nonlinear regression methods	542
13.4	Different estimates of the VCE	555
13.5	Prediction	562
13.6	Predictive margins	567
13.7	Marginal effects	570
13.8	Model diagnostics	585
13.9	Clustered data	589
13.10	Additional resources	596
13.11	Exercises	596
14	Flexible regression: finite mixtures and nonparametric	599
14.1	Introduction	599
14.2	Models based on finite mixtures	600
14.3	FMM example: Earnings of doctors	606
14.4	Global polynomials	618
14.5	Regression splines	621
14.6	Nonparametric regression	627
14.7	Partially parametric regression	632
14.8	Additional resources	632
14.9	Exercises	633
15	Quantile regression	635
15.1	Introduction	635
15.2	Conditional quantile regression	636
15.3	Conditional QR for medical expenditures data	639
15.4	Conditional QR for generated heteroskedastic data	650
15.5	Quantile treatment effects for a binary treatment	654
15.6	Additional resources	657
15.7	Exercises	657
16	Nonlinear optimization methods	661
16.1	Introduction	661
16.2	Newton–Raphson method	661

16.3	Gradient methods	666
16.4	Overview of ml, moptimize and optimize commands	670
16.5	The ml command: lf method	673
16.6	Checking the program	678
16.7	The ml command: lf0-lf2, d0-d2 and gf0 methods	684
16.8	Nonlinear IV (GMM) example	691
16.9	Additional resources	694
16.10	Exercises	694
17	Binary outcome models	697
17.1	Introduction	697
17.2	Some parametric models	697
17.3	Estimation	700
17.4	Example	702
17.5	Goodness of fit and prediction	708
17.6	Marginal effects	715
17.7	Clustered data	718
17.8	Additional models	720
17.9	Endogenous regressors	724
17.10	Grouped and aggregate data	732
17.11	Additional resources	735
17.12	Exercises	735
18	Multinomial models	737
18.1	Introduction	737
18.2	Multinomial models overview	737
18.3	Multinomial example: choice of fishing mode	741
18.4	Multinomial logit model	744
18.5	Alternative-specific conditional logit model	749
18.6	Nested logit model	757
18.7	Multinomial probit model	763
18.8	Alternative-specific random-parameters logit	768

18.9	Ordered outcome models	771
18.10	Clustered data	776
18.11	Multivariate outcomes	776
18.12	Additional resources	780
18.13	Exercises	780
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	846
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	870
20.5	Finite-mixture models	876
20.6	Zero-inflated models	893
20.7	Endogenous regressors	901
20.8	Clustered data	910
20.9	QR for count data	911
20.10	Additional resources	917

20.11	Exercises	917
21	Survival analysis for duration data	919
21.1	Introduction	919
21.2	Data and data summary	920
21.3	Survivor and hazard functions	924
21.4	Semiparametric regression model	929
21.5	Fully parametric regression models	937
21.6	Multiple-records data	947
21.7	Discrete-time hazards logit model	950
21.8	Time-varying regressors	953
21.9	Clustered data	954
21.10	Additional resources	954
21.11	Exercises	954
22	Nonlinear panel models	957
22.1	Introduction	957
22.2	Nonlinear panel-data overview	957
22.3	Nonlinear panel-data example	962
22.4	Binary outcome and ordered outcome models	965
22.5	Tobit and interval-data models	982
22.6	Count-data models	986
22.7	Panel quantile regression	997
22.8	Endogenous regressors in nonlinear panel models	1000
22.9	Additional resources	1000
22.10	Exercises	1000
23	Parametric models for heterogeneity and endogeneity	1003
23.1	Introduction	1003
23.2	Finite mixtures and unobserved heterogeneity	1004
23.3	Empirical examples of finite mixture models	1006
23.4	Nonlinear mixed effects models	1033
23.5	SEM for linear structural equation models	1039

23.6	Generalized SEM	1059
23.7	ERM commands for endogeneity and selection	1068
23.8	Additional resources	1073
23.9	Exercises	1073
24	RCTs and exogenous treatment effects	1075
24.1	Introduction	1075
24.2	Potential outcomes	1077
24.3	Randomized controlled trials	1078
24.4	Regression in an RCT	1087
24.5	Treatment evaluation with exogenous treatment	1095
24.6	Treatment evaluation methods and estimators	1097
24.7	Stata commands for treatment evaluation	1106
24.8	Oregon Health Insurance Experiment	1109
24.9	Treatment effect estimates using the OHIE data	1115
24.10	Multilevel treatment effects	1125
24.11	Conditional quantile treatment effects	1133
24.12	Additional resources	1135
24.13	Exercises	1135
25	Endogenous treatment effects	1137
25.1	Introduction	1137
25.2	Parametric methods for endogenous treatment	1138
25.3	ERM commands for endogenous treatment	1141
25.4	ET commands for binary endogenous treatment	1148
25.5	The LATE estimator for heterogeneous effects	1156
25.6	Differences-in-differences and synthetic control	1162
25.7	Regression discontinuity design	1167
25.8	Conditional QR with endogenous regressors	1185
25.9	Unconditional quantiles	1190
25.10	Additional resources	1197
25.11	Exercises	1198

26	Spatial regression	1201
26.1	Introduction	1201
26.2	Overview of spatial regression models	1201
26.3	Geospatial data	1203
26.4	The spatial weighting matrix	1206
26.5	OLS regression and test for spatial correlation	1209
26.6	Spatial dependence in the error	1210
26.7	Spatial autoregressive (SAR) models	1212
26.8	Spatial instrumental variables	1222
26.9	Spatial panel-data models	1223
26.10	Additional resources	1224
26.11	Exercises	1224
27	Semiparametric regression	1227
27.1	Introduction	1227
27.2	Kernel regression	1228
27.3	Series regression	1232
27.4	Nonparametric single regressor example	1233
27.5	Nonparametric multiple regressor example	1243
27.6	Partial linear model	1245
27.7	Single-index model	1248
27.8	Generalized additive model	1250
27.9	Additional resources	1253
27.10	Exercises	1253
28	Machine learning for prediction and inference	1255
28.1	Introduction	1255
28.2	Measuring the predictive ability of a model	1256
28.3	Shrinkage Estimators	1266
28.4	Prediction using LASSO, ridge and elasticnet	1271
28.5	Dimension reduction	1281
28.6	Machine learning methods for prediction	1284

28.7	Prediction application	1289
28.8	Machine learning for inference in partial linear model	1294
28.9	Machine learning for inference in other models	1303
28.10	Additional resources	1309
28.11	Exercises	1310
29	Bayesian methods: basics	1313
29.1	Introduction	1313
29.2	Bayesian introductory example	1314
29.3	Bayesian methods overview	1318
29.4	An i.i.d. example	1324
29.5	Linear regression	1335
29.6	A linear regression example	1338
29.7	Modifying the MH algorithm	1344
29.8	Random-effects model	1347
29.9	Bayesian model selection	1351
29.10	Bayesian prediction	1354
29.11	Probit example	1357
29.12	Additional resources	1361
29.13	Exercises	1361
30	Bayesian methods: MCMC Algorithms	1363
30.1	Introduction	1363
30.2	User-provided log likelihood	1363
30.3	MH algorithm in Mata	1368
30.4	Data augmentation and the Gibbs sampler in Mata	1373
30.5	Multiple imputation	1378
30.6	Multiple imputation example	1383
30.7	Regression with complete and incomplete data	1383
30.8	Additional resources	1392
30.9	Exercises	1392
A	Programming in Stata	1401

A.1	Stata matrix commands	1401
A.2	Programs	1407
A.3	Program debugging	1413
A.4	Additional resources	1416
B	Mata	1417
B.1	How to run Mata	1417
B.2	Mata matrix commands	1419
B.3	Programming in Mata	1428
B.4	Additional resources	1430
C	Optimization in Mata	1431
C.1	Mata moptimize function	1431
C.2	Mata optimize() function	1440
C.3	Additional resources	1444
	References	1445
	Author index	1469
	Subject index	1475