

# Microeconometrics Using Stata

Second Edition

DRAFT TO STATA PRESS FOR PRODUCTION

NOVEMBER 2020

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	xvii
	List of figures	xix
	Preface to the Second Edition	xxiii
	Preface to the First Edition	xxv
<b>1</b>	<b>Stata basics</b>	<b>1</b>
	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
<b>2</b>	<b>Data management and graphics</b>	<b>31</b>
	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
<b>3</b>	<b>Linear regression basics</b>	<b>81</b>
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 . . . . .	123
3.8	Sampling weights . . . . .	131
3.9	OLS using Mata . . . . .	135
3.10	Additional resources . . . . .	137
3.11	Exercises . . . . .	137
<b>4</b>	<b>Linear regression extensions</b>	<b>139</b>
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 . . . . .	180
4.8	Differences-in-differences estimators . . . . .	181
4.9	Additional resources . . . . .	188
4.10	Exercises . . . . .	188
<b>5</b>	<b>Simulation</b>	<b>191</b>
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
<b>6</b>	<b>Linear regression with correlated errors</b>	<b>227</b>
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 . . . . .	257
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
<b>7</b>	<b>Linear instrumental variables regression</b>	<b>283</b>
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 . . . . .	342
7.11	Additional resources . . . . .	343

7.12	Exercises . . . . .	344
<b>8</b>	<b>Linear panel-data models: basics</b>	<b>347</b>
8.1	Introduction . . . . .	347
8.2	Panel-data methods overview . . . . .	347
8.3	Summary of panel-data . . . . .	352
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
<b>9</b>	<b>Linear panel-data models: extensions</b>	<b>393</b>
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
<b>10</b>	<b>Introduction to nonlinear regression</b>	<b>427</b>
10.1	Introduction . . . . .	427
10.2	Binary outcome models . . . . .	427
10.3	Probit model . . . . .	430
10.4	Marginal effects . . . . .	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
<b>11</b>	<b>Tests of hypotheses and model specification</b>	<b>445</b>
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 . . . . .	497
<b>12</b>	<b>Bootstrap methods</b>	<b>499</b>
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 . . . . .	537
<b>13</b>	<b>Nonlinear regression methods</b>	<b>539</b>
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 . . . . .	595
13.11	Exercises . . . . .	596
<b>14</b>	<b>Flexible regression: finite mixtures and nonparametric</b>	<b>599</b>
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 . . . . .	633
14.9	Exercises . . . . .	633
<b>15</b>	<b>Quantile regression</b>	<b>635</b>
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 . . . . .	651
15.5	Quantile treatment effects for a binary treatment . . . . .	654
15.6	Additional resources . . . . .	657
15.7	Exercises . . . . .	657
<b>16</b>	<b>Nonlinear optimization methods</b>	<b>661</b>
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 . . . . .	672
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
<b>17</b>	<b>Binary outcome models</b>	<b>697</b>
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 . . . . .	719
17.9	Endogenous regressors . . . . .	724
17.10	Grouped and aggregate data . . . . .	732
17.11	Additional resources . . . . .	735
17.12	Exercises . . . . .	735
<b>18</b>	<b>Multinomial models</b>	<b>737</b>
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 . . . . .	775
18.11	Multivariate outcomes . . . . .	776
18.12	Additional resources . . . . .	779
18.13	Exercises . . . . .	780
<b>19</b>	<b>Tobit and selection models</b>	<b>781</b>
19.1	Introduction . . . . .	781
19.2	Tobit model . . . . .	782
19.3	Tobit model example . . . . .	784
19.4	Tobit for lognormal data . . . . .	793
19.5	Two-part model in logs . . . . .	801
19.6	Selection models . . . . .	804
19.7	Non-normal models of selection . . . . .	811
19.8	Prediction from models with outcome in logs . . . . .	815
19.9	Endogenous regressors . . . . .	818
19.10	Missing data . . . . .	819
19.11	Panel attrition . . . . .	824
19.12	Additional resources . . . . .	845
19.13	Exercises . . . . .	845
<b>20</b>	<b>Count-data models</b>	<b>847</b>
20.1	Introduction . . . . .	847
20.2	Modeling strategies for count data . . . . .	848
20.3	Poisson and negative binomial models . . . . .	852
20.4	Hurdle model . . . . .	867
20.5	Finite-mixture models . . . . .	873
20.6	Zero-inflated models . . . . .	891
20.7	Endogenous regressors . . . . .	899
20.8	Clustered data . . . . .	908
20.9	QR for count data . . . . .	910
20.10	Additional resources . . . . .	915

20.11 Exercises . . . . .	916
<b>21 Survival analysis for duration data</b>	<b>919</b>
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 . . . . .	949
21.8 Time-varying regressors . . . . .	953
21.9 Clustered data . . . . .	953
21.10 Additional resources . . . . .	954
21.11 Exercises . . . . .	954
<b>22 Nonlinear panel models</b>	<b>957</b>
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 . . . . .	999
22.9 Additional resources . . . . .	1000
22.10 Exercises . . . . .	1000
<b>23 Parametric models for heterogeneity and endogeneity</b>	<b>1003</b>
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 . . . . .	1032
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 . . . . .	1072
23.9	Exercises . . . . .	1073
<b>24</b>	<b>RCTs and exogenous treatment effects</b>	<b>1075</b>
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 . . . . .	1107
24.8	Oregon Health Insurance Experiment . . . . .	1109
24.9	Treatment effect estimates using the OHIE data . . . . .	1116
24.10	Multilevel treatment effects . . . . .	1125
24.11	Conditional quantile treatment effects . . . . .	1133
24.12	Additional resources . . . . .	1135
24.13	Exercises . . . . .	1136
<b>25</b>	<b>Endogenous treatment effects</b>	<b>1139</b>
25.1	Introduction . . . . .	1139
25.2	Parametric methods for endogenous treatment . . . . .	1140
25.3	ERM commands for endogenous treatment . . . . .	1143
25.4	ET commands for binary endogenous treatment . . . . .	1150
25.5	The LATE estimator for heterogeneous effects . . . . .	1158
25.6	Differences-in-differences and synthetic control . . . . .	1164
25.7	Regression discontinuity design . . . . .	1168
25.8	Conditional QR with endogenous regressors . . . . .	1186
25.9	Unconditional quantiles . . . . .	1192
25.10	Additional resources . . . . .	1198
25.11	Exercises . . . . .	1199

<b>26</b>	<b>Spatial regression</b>	<b>1201</b>
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
<b>27</b>	<b>Semiparametric regression</b>	<b>1227</b>
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 . . . . .	1249
27.8	Generalized additive model . . . . .	1251
27.9	Additional resources . . . . .	1253
27.10	Exercises . . . . .	1254
<b>28</b>	<b>Machine learning for prediction and inference</b>	<b>1257</b>
28.1	Introduction . . . . .	1257
28.2	Measuring the predictive ability of a model . . . . .	1258
28.3	Shrinkage Estimators . . . . .	1268
28.4	Prediction using LASSO, ridge and elasticnet . . . . .	1273
28.5	Dimension reduction . . . . .	1283
28.6	Machine learning methods for prediction . . . . .	1286

28.7	Prediction application . . . . .	1291
28.8	Machine learning for inference in partial linear model . . . . .	1296
28.9	Machine learning for inference in other models . . . . .	1304
28.10	Additional resources . . . . .	1310
28.11	Exercises . . . . .	1311
<b>29</b>	<b>Bayesian methods: basics</b>	<b>1313</b>
29.1	Introduction . . . . .	1313
29.2	Bayesian introductory example . . . . .	1314
29.3	Bayesian methods overview . . . . .	1317
29.4	An i.i.d. example . . . . .	1323
29.5	Linear regression . . . . .	1333
29.6	A linear regression example . . . . .	1336
29.7	Modifying the MH algorithm . . . . .	1343
29.8	Random effects model . . . . .	1346
29.9	Bayesian model selection . . . . .	1349
29.10	Bayesian prediction . . . . .	1351
29.11	Probit example . . . . .	1354
29.12	Additional resources . . . . .	1358
29.13	Exercises . . . . .	1358
<b>30</b>	<b>Bayesian methods: MCMC Algorithms</b>	<b>1361</b>
30.1	Introduction . . . . .	1361
30.2	User-provided log-likelihood . . . . .	1361
30.3	Metropolis-Hastings algorithm in Mata . . . . .	1365
30.4	Data augmentation and the Gibbs sampler in Mata . . . . .	1371
30.5	Multiple imputation . . . . .	1376
30.6	Multiple imputation example . . . . .	1379
30.7	Regression with complete and incomplete data . . . . .	1380
30.8	Additional resources . . . . .	1388
30.9	Exercises . . . . .	1388
<b>A</b>	<b>Programming in Stata</b>	<b>1397</b>

A.1	Stata matrix commands . . . . .	1397
A.2	Programs . . . . .	1403
A.3	Program debugging . . . . .	1409
A.4	Additional resources . . . . .	1412
<b>B</b>	<b>Mata</b>	<b>1413</b>
B.1	How to run Mata . . . . .	1413
B.2	Mata matrix commands . . . . .	1415
B.3	Programming in Mata . . . . .	1424
B.4	Additional resources . . . . .	1426
<b>C</b>	<b>Optimization in Mata</b>	<b>1427</b>
C.1	Mata moptimize function . . . . .	1427
C.2	Mata optimize() function . . . . .	1436
C.3	Additional resources . . . . .	1440
	<b>References</b>	<b>1441</b>
	<b>Author index</b>	<b>1465</b>
	<b>Subject index</b>	<b>1471</b>