

A 334 961 484 ✓

Data Analysis Using Stata

Third Edition

Contents

	List of tables	xvii
	List of figures	xix
	Preface	xxi
	Acknowledgments	xxvii
1	The first time	1
1.1	Starting Stata	1
1.2	Setting up your screen	2
1.3	Your first analysis	2
1.3.1	Inputting commands	2
1.3.2	Files and the working memory	3
1.3.3	Loading data	3
1.3.4	Variables and observations	5
1.3.5	Looking at data	7
1.3.6	Interrupting a command and repeating a command	8
1.3.7	The variable list	8
1.3.8	The in qualifier	9
1.3.9	Summary statistics	9
1.3.10	The if qualifier	11
1.3.11	Defining missing values	11
1.3.12	The by prefix	12
1.3.13	Command options	13
1.3.14	Frequency tables	14
1.3.15	Graphs	15
1.3.16	Getting help	16

1.3.17	Recoding variables	17
1.3.18	Variable labels and value labels	18
1.3.19	Linear regression	19
1.4	Do-files	20
1.5	Exiting Stata	22
1.6	Exercises	23
2	Working with do-files	25
2.1	From interactive work to working with a do-file	25
2.1.1	Alternative 1	26
2.1.2	Alternative 2	27
2.2	Designing do-files	30
2.2.1	Comments	31
2.2.2	Line breaks	32
2.2.3	Some crucial commands	33
2.3	Organizing your work	35
2.4	Exercises	39
3	The grammar of Stata	41
3.1	The elements of Stata commands	41
3.1.1	Stata commands	41
3.1.2	The variable list	43
List of variables: Required or optional	43	
Abbreviation rules	43	
Special listings	45	
3.1.3	Options	45
3.1.4	The in qualifier	47
3.1.5	The if qualifier	48
3.1.6	Expressions	51
Operators	52	
Functions	54	
3.1.7	Lists of numbers	55

3.1.8	Using filenames	56
3.2	Repeating similar commands	57
3.2.1	The by prefix	58
3.2.2	The foreach loop	59
	The types of foreach lists	61
	Several commands within a foreach loop	62
3.2.3	The forvalues loop	62
3.3	Weights	63
	Frequency weights	64
	Analytic weights	66
	Sampling weights	67
3.4	Exercises	68
4	General comments on the statistical commands	71
4.1	Regular statistical commands	71
4.2	Estimation commands	74
4.3	Exercises	76
5	Creating and changing variables	77
5.1	The commands generate and replace	77
5.1.1	Variable names	78
5.1.2	Some examples	79
5.1.3	Useful functions	82
5.1.4	Changing codes with by, _n, and _N	85
5.1.5	Subscripts	89
5.2	Specialized recoding commands	91
5.2.1	The recode command	91
5.2.2	The egen command	92
5.3	Recoding string variables	94
5.4	Recoding date and time	98
5.4.1	Dates	98
5.4.2	Time	102

5.5	Setting missing values	105
5.6	Labels	107
5.7	Storage types, or the ghost in the machine	111
5.8	Exercises	112
6	Creating and changing graphs	115
6.1	A primer on graph syntax	115
6.2	Graph types	116
6.2.1	Examples	117
6.2.2	Specialized graphs	119
6.3	Graph elements	119
6.3.1	Appearance of data	121
	Choice of marker	123
	Marker colors	125
	Marker size	126
	Lines	126
6.3.2	Graph and plot regions	129
	Graph size	130
	Plot region	130
	Scaling the axes	131
6.3.3	Information inside the plot region	133
	Reference lines	133
	Labeling inside the plot region	134
6.3.4	Information outside the plot region	138
	Labeling the axes	139
	Tick lines	142
	Axis titles	143
	The legend	144
	Graph titles	146
6.4	Multiple graphs	147
6.4.1	Overlaying many twoway graphs	147

6.4.2	Option by()	149
6.4.3	Combining graphs	150
6.5	Saving and printing graphs	152
6.6	Exercises	154
7	Describing and comparing distributions	157
7.1	Categories: Few or many?	158
7.2	Variables with few categories	159
7.2.1	Tables	159
	Frequency tables	159
	More than one frequency table	160
	Comparing distributions	160
	Summary statistics	162
	More than one contingency table	163
7.2.2	Graphs	163
	Histograms	164
	Bar charts	166
	Pie charts	168
	Dot charts	169
7.3	Variables with many categories	170
7.3.1	Frequencies of grouped data	171
	Some remarks on grouping data	171
	Special techniques for grouping data	172
7.3.2	Describing data using statistics	173
	Important summary statistics	174
	The summarize command	176
	The tabstat command	177
	Comparing distributions using statistics	178
7.3.3	Graphs	186
	Box plots	187
	Histograms	189

	Kernel density estimation	191
	Quantile plot	195
	Comparing distributions with Q–Q plots	199
7.4	Exercises	200
8	Statistical inference	201
8.1	Random samples and sampling distributions	202
8.1.1	Random numbers	202
8.1.2	Creating fictitious datasets	203
8.1.3	Drawing random samples	207
8.1.4	The sampling distribution	208
8.2	Descriptive inference	213
8.2.1	Standard errors for simple random samples	213
8.2.2	Standard errors for complex samples	215
	Typical forms of complex samples	215
	Sampling distributions for complex samples	217
	Using Stata's svy commands	219
8.2.3	Standard errors with nonresponse	222
	Unit nonresponse and poststratification weights	222
	Item nonresponse and multiple imputation	223
8.2.4	Uses of standard errors	230
	Confidence intervals	231
	Significance tests	233
	Two-group mean comparison test	238
8.3	Causal inference	242
8.3.1	Basic concepts	242
	Data-generating processes	242
	Counterfactual concept of causality	244
8.3.2	The effect of third-class tickets	246
8.3.3	Some problems of causal inference	248
8.4	Exercises	250

9	Introduction to linear regression	253
9.1	Simple linear regression	256
9.1.1	The basic principle	256
9.1.2	Linear regression using Stata	260
	The table of coefficients	261
	The table of ANOVA results	266
	The model fit table	268
9.2	Multiple regression	270
9.2.1	Multiple regression using Stata	271
9.2.2	More computations	274
	Adjusted R^2	274
	Standardized regression coefficients	276
9.2.3	What does “under control” mean?	277
9.3	Regression diagnostics	279
9.3.1	Violation of $E(\epsilon_i) = 0$	280
	Linearity	283
	Influential cases	286
	Omitted variables	295
	Multicollinearity	296
9.3.2	Violation of $\text{Var}(\epsilon_i) = \sigma^2$	296
9.3.3	Violation of $\text{Cov}(\epsilon_i, \epsilon_j) = 0, i \neq j$	299
9.4	Model extensions	300
9.4.1	Categorical independent variables	301
9.4.2	Interaction terms	304
9.4.3	Regression models using transformed variables	308
	Nonlinear relationships	309
	Eliminating heteroskedasticity	312
9.5	Reporting regression results	313
9.5.1	Tables of similar regression models	313
9.5.2	Plots of coefficients	316

9.5.3	Conditional-effects plots	321
9.6	Advanced techniques	324
9.6.1	Median regression	324
9.6.2	Regression models for panel data	327
	From wide to long format	328
	Fixed-effects models	332
9.6.3	Error-components models	337
9.7	Exercises	339
10	Regression models for categorical dependent variables	341
10.1	The linear probability model	342
10.2	Basic concepts	346
10.2.1	Odds, log odds, and odds ratios	346
10.2.2	Excursion: The maximum likelihood principle	351
10.3	Logistic regression with Stata	354
10.3.1	The coefficient table	356
	Sign interpretation	357
	Interpretation with odds ratios	357
	Probability interpretation	359
	Average marginal effects	361
10.3.2	The iteration block	362
10.3.3	The model fit block	363
	Classification tables	364
	Pearson chi-squared	367
10.4	Logistic regression diagnostics	368
10.4.1	Linearity	369
10.4.2	Influential cases	372
10.5	Likelihood-ratio test	377
10.6	Refined models	379
10.6.1	Nonlinear relationships	379
10.6.2	Interaction effects	381

10.7	Advanced techniques	384
10.7.1	Probit models	385
10.7.2	Multinomial logistic regression	387
10.7.3	Models for ordinal data	391
10.8	Exercises	393
11	Reading and writing data	395
11.1	The goal: The data matrix	395
11.2	Importing machine-readable data	397
11.2.1	Reading system files from other packages	398
	Reading Excel files	398
	Reading SAS transport files	402
	Reading other system files	402
11.2.2	Reading ASCII text files	402
	Reading data in spreadsheet format	402
	Reading data in free format	405
	Reading data in fixed format	407
11.3	Inputting data	410
11.3.1	Input data using the Data Editor	410
11.3.2	The input command	411
11.4	Combining data	415
11.4.1	The GSOEP database	415
11.4.2	The merge command	417
	Merge 1:1 matches with rectangular data	418
	Merge 1:1 matches with nonrectangular data	421
	Merging more than two files	424
	Merging m:1 and 1:m matches	425
11.4.3	The append command	429
11.5	Saving and exporting data	433
11.6	Handling large datasets	434
11.6.1	Rules for handling the working memory	434

11.6.2	Using oversized datasets	435
11.7	Exercises	435
12	Do-files for advanced users and user-written programs	437
12.1	Two examples of usage	437
12.2	Four programming tools	439
12.2.1	Local macros	439
	Calculating with local macros	440
	Combining local macros	441
	Changing local macros	442
12.2.2	Do-files	443
12.2.3	Programs	443
	The problem of redefinition	445
	The problem of naming	445
	The problem of error checking	445
12.2.4	Programs in do-files and ado-files	446
12.3	User-written Stata commands	449
12.3.1	Sketch of the syntax	451
12.3.2	Create a first ado-file	452
12.3.3	Parsing variable lists	453
12.3.4	Parsing options	454
12.3.5	Parsing if and in qualifiers	456
12.3.6	Generating an unknown number of variables	457
12.3.7	Default values	459
12.3.8	Extended macro functions	461
12.3.9	Avoiding changes in the dataset	463
12.3.10	Help files	465
12.4	Exercises	467
13	Around Stata	469
13.1	Resources and information	469
13.2	Taking care of Stata	470

13.3	Additional procedures	471
13.3.1	Stata Journal ado-files	471
13.3.2	SSC ado-files	473
13.3.3	Other ado-files	474
13.4	Exercises	475
	References	477
	Author index	483
	Subject index	487