

Contents

List of Figures	xiii
List of Tables	xv
Preface	xvii
1 Data management	1
1.1 Input	1
1.1.1 Native dataset	1
1.1.2 Fixed format text files	2
1.1.3 Reading more complex text files	3
1.1.4 Comma separated value (CSV) files	4
1.1.5 Reading datasets in other formats	4
1.1.6 URL	5
1.1.7 XML (extensible markup language)	6
1.1.8 Data entry	7
1.2 Output	7
1.2.1 Save a native dataset	7
1.2.2 Creating files for use by other packages	8
1.2.3 Creating datasets in text format	9
1.2.4 Displaying data	9
1.2.5 Number of digits to display	10
1.2.6 Creating HTML formatted output	10
1.2.7 Creating XML datasets and output	11
1.3 Structure and meta-data	11
1.3.1 Access variables from a dataset	11
1.3.2 Names of variables and their types	12
1.3.3 Values of variables in a dataset	12
1.3.4 Rename variables in a dataset	12
1.3.5 Add comment to a dataset or variable	13
1.4 Derived variables and data manipulation	13
1.4.1 Create string variables from numeric variables	13
1.4.2 Create numeric variables from string variables	14
1.4.3 Extract characters from string variables	14
1.4.4 Length of string variables	15
1.4.5 Concatenate string variables	15
1.4.6 Find strings within string variables	15
1.4.7 Remove spaces around string variables	16
1.4.8 Upper to lower case	16

1.4.9	Create categorical variables from continuous variables	17
1.4.10	Recode a categorical variable	17
1.4.11	Create a categorical variable using logic	18
1.4.12	Formatting values of variables	18
1.4.13	Label variables	19
1.4.14	Account for missing values	19
1.4.15	Observation number	21
1.4.16	Unique values	22
1.4.17	Lagged variable	22
1.4.18	SQL	23
1.4.19	Perl interface	23
1.5	Merging, combining, and subsetting datasets	23
1.5.1	Subsetting observations	23
1.5.2	Random sample of a dataset	24
1.5.3	Convert from wide to long (tall) format	25
1.5.4	Convert from long (tall) to wide format	26
1.5.5	Concatenate datasets	26
1.5.6	Sort datasets	27
1.5.7	Merge datasets	27
1.5.8	Drop variables in a dataset	29
1.6	Date and time variables	30
1.6.1	Create date variable	30
1.6.2	Extract weekday	30
1.6.3	Extract month	31
1.6.4	Extract year	31
1.6.5	Extract quarter	31
1.6.6	Create time variable	31
1.7	Interactions with the operating system	32
1.7.1	Timing commands	32
1.7.2	Execute command in operating system	32
1.7.3	Find working directory	33
1.7.4	Change working directory	33
1.7.5	List and access files	34
1.8	Mathematical functions	34
1.8.1	Basic functions	34
1.8.2	Trigonometric functions	35
1.8.3	Special functions	35
1.8.4	Integer functions	36
1.8.5	Comparisons of floating point variables	36
1.8.6	Derivative	37
1.8.7	Optimization problems	37
1.9	Matrix operations	38
1.9.1	Create matrix	38
1.9.2	Transpose matrix	38
1.9.3	Invert matrix	39
1.9.4	Create submatrix	39
1.9.5	Create a diagonal matrix	39
1.9.6	Create vector of diagonal elements	40
1.9.7	Create vector from a matrix	40
1.9.8	Calculate determinant	40
1.9.9	Find eigenvalues and eigenvectors	40

1.9.10	Calculate singular value decomposition	41
1.10	Probability distributions and random number generation	41
1.10.1	Probability density function	41
1.10.2	Quantiles of a probability density function	42
1.10.3	Uniform random variables	42
1.10.4	Multinomial random variables	42
1.10.5	Normal random variables	44
1.10.6	Multivariate normal random variables	44
1.10.7	Exponential random variables	45
1.10.8	Other random variables	46
1.10.9	Setting the random number seed	46
1.11	Control flow, programming, and data generation	47
1.11.1	Looping	47
1.11.2	Conditional execution	47
1.11.3	Sequence of values or patterns	48
1.11.4	Referring to a range of variables	50
1.11.5	Perform an action repeatedly over a set of variables	50
1.12	Further resources	51
1.13	HELP examples	51
1.13.1	Data input and output	51
1.13.2	Data display	54
1.13.3	Derived variables and data manipulation	55
1.13.4	Sorting and subsetting datasets	61
1.13.5	Probability distributions	63
2	Common statistical procedures	65
2.1	Summary statistics	65
2.1.1	Means and other summary statistics	65
2.1.2	Means by group	66
2.1.3	Trimmed mean	67
2.1.4	Five-number summary	67
2.1.5	Quantiles	67
2.1.6	Centering, normalizing, and scaling	68
2.1.7	Mean and 95% confidence interval	68
2.1.8	Bootstrapping a sample statistic	69
2.1.9	Proportion and 95% confidence interval	70
2.2	Bivariate statistics	70
2.2.1	Epidemiologic statistics	70
2.2.2	Test characteristics	71
2.2.3	Correlation	72
2.2.4	Kappa (agreement)	73
2.3	Contingency tables	73
2.3.1	Display cross-classification table	73
2.3.2	Pearson chi-square statistic	74
2.3.3	Cochran–Mantel–Haenszel test	74
2.3.4	Fisher’s exact test	75
2.3.5	McNemar’s test	75
2.4	Two sample tests for continuous variables	75
2.4.1	Student’s t-test	75
2.4.2	Nonparametric tests	76
2.4.3	Permutation test	76

2.4.4	Logrank test	77
2.5	Further resources	77
2.6	HELP examples	78
2.6.1	Summary statistics and exploratory data analysis	78
2.6.2	Bivariate relationships	80
2.6.3	Contingency tables	82
2.6.4	Two sample tests of continuous variables	85
2.6.5	Survival analysis: logrank test	90
3	Linear regression and ANOVA	93
3.1	Model fitting	93
3.1.1	Linear regression	93
3.1.2	Linear regression with categorical covariates	94
3.1.3	Parameterization of categorical covariates	94
3.1.4	Linear regression with no intercept	96
3.1.5	Linear regression with interactions	96
3.1.6	Linear models stratified by each value of a grouping variable	97
3.1.7	One-way analysis of variance	97
3.1.8	Two-way (or more) analysis of variance	98
3.2	Model comparison and selection	98
3.2.1	Compare two models	98
3.2.2	Log-likelihood	99
3.2.3	Akaike Information Criterion (AIC)	99
3.2.4	Bayesian Information Criterion (BIC)	99
3.3	Tests, contrasts, and linear functions of parameters	100
3.3.1	Joint null hypotheses: several parameters equal 0	100
3.3.2	Joint null hypotheses: sum of parameters	100
3.3.3	Tests of equality of parameters	101
3.3.4	Multiple comparisons	101
3.3.5	Linear combinations of parameters	102
3.4	Model diagnostics	102
3.4.1	Predicted values	102
3.4.2	Residuals	103
3.4.3	Studentized residuals	103
3.4.4	Leverage	104
3.4.5	Cook's D	104
3.4.6	DFFITs	105
3.4.7	Diagnostic plots	106
3.5	Model parameters and results	106
3.5.1	Prediction limits	106
3.5.2	Parameter estimates	107
3.5.3	Standard errors of parameter estimates	107
3.5.4	Confidence limits for the mean	108
3.5.5	Plot confidence intervals for the mean	108
3.5.6	Plot prediction limits from a simple linear regression	109
3.5.7	Plot predicted lines for each value of a variable	109
3.5.8	Design and information matrix	110
3.5.9	Covariance matrix	110
3.6	Further resources	111
3.7	HELP examples	111
3.7.1	Scatterplot with smooth fit	111

3.7.2	Linear regression with interaction	113
3.7.3	Regression diagnostics	116
3.7.4	Fitting regression model separately for each value of another variable	119
3.7.5	Two way ANOVA	120
3.7.6	Multiple comparisons	126
3.7.7	Contrasts	128
4	Regression generalizations	131
4.1	Generalized linear models	131
4.1.1	Logistic regression model	131
4.1.2	Exact logistic regression	133
4.1.3	Poisson model	134
4.1.4	Zero-inflated Poisson model	134
4.1.5	Negative binomial model	135
4.1.6	Zero-inflated negative binomial model	135
4.1.7	Log-linear model	136
4.1.8	Ordered multinomial model	136
4.1.9	Generalized (nominal outcome) multinomial logit	137
4.1.10	Conditional logistic regression model	137
4.2	Models for correlated data	137
4.2.1	Linear models with correlated outcomes	137
4.2.2	Linear mixed models with random intercepts	138
4.2.3	Linear mixed models with random slopes	139
4.2.4	More complex random coefficient models	140
4.2.5	Multilevel models	140
4.2.6	Generalized linear mixed models	141
4.2.7	Generalized estimating equations	141
4.2.8	Time series model	142
4.3	Survival analysis	143
4.3.1	Proportional hazards (Cox) regression model	143
4.3.2	Proportional hazards (Cox) model with frailty	143
4.4	Further generalizations to regression models	143
4.4.1	Nonlinear least squares model	143
4.4.2	Generalized additive model	144
4.4.3	Robust regression model	144
4.4.4	Quantile regression model	145
4.4.5	Ridge regression model	145
4.5	Further resources	146
4.6	HELP examples	146
4.6.1	Logistic regression	146
4.6.2	Poisson regression	150
4.6.3	Zero-inflated Poisson regression	152
4.6.4	Negative binomial regression	154
4.6.5	Quantile regression	155
4.6.6	Ordinal logit	156
4.6.7	Multinomial logit	157
4.6.8	Generalized additive model	159
4.6.9	Reshaping dataset for longitudinal regression	160
4.6.10	Linear model for correlated data	164
4.6.11	Linear mixed (random slope) model	166
4.6.12	Generalized estimating equations	171

4.6.13	Generalized linear mixed model	172
4.6.14	Cox proportional hazards model	173
5	Graphics	177
5.1	A compendium of useful plots	178
5.1.1	Scatterplot	178
5.1.2	Scatterplot with multiple y values	178
5.1.3	Barplot	179
5.1.4	Histogram	180
5.1.5	Stem-and-leaf plot	181
5.1.6	Boxplot	181
5.1.7	Side-by-side boxplots	182
5.1.8	Normal quantile-quantile plot	182
5.1.9	Interaction plots	183
5.1.10	Plots for categorical data	183
5.1.11	Conditioning plot	184
5.1.12	3-D plots	184
5.1.13	Circular plot	185
5.1.14	Sunflower plot	185
5.1.15	Empirical cumulative probability density plot	185
5.1.16	Empirical probability density plot	186
5.1.17	Matrix of scatterplots	186
5.1.18	Receiver operating characteristic (ROC) curve	187
5.1.19	Kaplan–Meier plot	187
5.2	Adding elements	188
5.2.1	Arbitrary straight line	189
5.2.2	Plot symbols	189
5.2.3	Add points to an existing graphic	190
5.2.4	Jitter	191
5.2.5	OLS line fit to points	191
5.2.6	Smoothed line	192
5.2.7	Normal density	192
5.2.8	Marginal rug plot	193
5.2.9	Titles	193
5.2.10	Footnotes	193
5.2.11	Text	194
5.2.12	Mathematical symbols	195
5.2.13	Arrows and shapes	195
5.2.14	Legend	196
5.2.15	Identifying and locating points	196
5.3	Options and parameters	197
5.3.1	Graph size	197
5.3.2	Point and text size	197
5.3.3	Box around plots	198
5.3.4	Size of margins	198
5.3.5	Graphical settings	198
5.3.6	Multiple plots per page	199
5.3.7	Axis range and style	199
5.3.8	Axis labels, values and tick marks	200
5.3.9	Line styles	200
5.3.10	Line widths	201

5.3.11	Colors	201
5.3.12	Log scale	201
5.3.13	Omit axes	202
5.4	Saving graphs	202
5.4.1	PDF	202
5.4.2	Postscript	203
5.4.3	RTF	203
5.4.4	JPEG	204
5.4.5	WMF	204
5.4.6	BMP	205
5.4.7	TIFF	205
5.4.8	PNG	206
5.4.9	Closing a graphic device	206
5.5	Further resources	206
5.6	HELP examples	206
5.6.1	Scatterplot with multiple axes	207
5.6.2	Conditioning plot	208
5.6.3	Kaplan–Meier plot	209
5.6.4	ROC curve	211
5.6.5	Pairs plot	213
5.6.6	Visualize correlation matrix	214
6	Other topics and extended examples	217
6.1	Power and sample size calculations	217
6.1.1	Analytic power calculation	217
6.1.2	Simulation-based power calculations	219
6.2	Generate data from generalized linear random effects model	222
6.3	Generate correlated binary data	223
6.4	Read variable format files and plot maps	224
6.4.1	Read input files	224
6.4.2	Plotting maps	226
6.5	Missing data: multiple imputation	228
6.6	Bayesian Poisson regression	231
6.7	Multivariate statistics and discriminant procedures	233
6.7.1	Cronbach’s α	233
6.7.2	Factor analysis	234
6.7.3	Recursive partitioning	237
6.7.4	Linear discriminant analysis	238
6.7.5	Hierarchical clustering	240
6.8	Complex survey design	241
6.9	Further resources	242
Appendix A	Introduction to SAS	243
A.1	Installation	243
A.2	Running SAS and a sample session	243
A.3	Learning SAS and getting help	247
A.4	Fundamental structures: data step, procedures, and global statements	249
A.5	Work process: The cognitive style of SAS	251
A.6	Useful SAS background	251
A.6.1	Data set options	251
A.6.2	Repeating commands for subgroups	252

A.6.3	Subsetting	252
A.6.4	Formats and informats	253
A.7	Accessing and controlling SAS output: the Output Delivery System	253
A.7.1	Saving output as datasets and controlling output	254
A.7.2	Output file types and ODS destinations	257
A.7.3	ODS graphics	257
A.8	The SAS Macro Facility: writing functions and passing values	258
A.8.1	Writing functions	258
A.8.2	SAS macro variables	258
A.9	Miscellanea	259
Appendix B Introduction to R		261
B.1	Installation	261
B.1.1	Installation under Windows	262
B.1.2	Installation under Mac OS X	262
B.1.3	Installation under Linux	262
B.2	Running R and sample session	263
B.2.1	Replicating examples from the book and sourcing commands	265
B.2.2	Batch mode	265
B.3	Learning R and getting help	265
B.4	Fundamental structures: objects, classes, and related concepts	266
B.4.1	Objects and vectors	266
B.4.2	Indexing	268
B.4.3	Operators	268
B.4.4	Matrices	268
B.4.5	Dataframes	269
B.4.6	Attributes and classes	271
B.5	Built-in and user-defined functions	271
B.5.1	Calling functions	271
B.5.2	Writing functions	272
B.5.3	The <code>apply</code> family of functions	273
B.6	Add-ons: libraries and packages	273
B.6.1	Introduction to libraries and packages	273
B.6.2	CRAN task views	274
B.6.3	Installed libraries and packages	274
B.6.4	Packages referenced in this book	275
B.6.5	Datasets available with R	276
B.7	Support and bugs	276
Appendix C The HELP study dataset		277
C.1	Background on the HELP study	277
C.2	Roadmap to analyses of the HELP dataset	277
C.3	Detailed description of the dataset	278
Appendix D References		283
Appendix E Indices		289
	Subject index	289
	SAS index	304
	R index	315

List of Figures

1.1	Comparison of standard normal and t distribution with 1 df	64
2.1	Density plot of depressive symptom scores (CESD) plus superimposed histogram and normal distribution	80
2.2	Scatterplot of CESD and MCS for women, with primary substance shown as the plot symbol	82
2.3	Density plot of age by gender	90
3.1	Scatterplot of observed values for AGE and I1 (plus smoothers by substance)	112
3.2	Q-Q plot from SAS, default diagnostics from R	118
3.3	Empirical density of residuals, with superimposed normal density	119
3.4	Interaction plot of CESD as a function of substance group and gender	121
3.5	Boxplot of CESD as a function of substance group and gender	122
3.6	Pairwise comparisons	128
4.1	Scatterplots of smoothed association of PCS with CESD	161
4.2	Side-by-side box plots of CESD by treatment and time	167
5.1	Plot of InDUC and MCS vs. CESD for female alcohol-involved subjects	208
5.2	Association of MCS and CESD, stratified by substance and report of suicidal thoughts	210
5.3	Kaplan–Meier estimate of time to linkage to primary care by randomization group	211
5.4	Receiver operating characteristic curve for the logistical regression model predicting suicidal thoughts using the CESD as a measure of depressive symptoms (sensitivity = true positive rate; 1-specificity = false positive rate)	212
5.5	Pairsplot of variables from the HELP dataset	214
5.6	Visual display of correlations and associations	216
6.1	Massachusetts counties	227
6.2	Recursive partitioning tree	238
6.3	Graphical display of assignment probabilities or score functions from linear discriminant analysis by actual homeless status	240
6.4	Results from hierarchical clustering	241
A.1	SAS Windows interface	244
A.2	Running a SAS program	245
A.3	The SAS window after running the sample session code	248
A.4	The SAS Explorer window	248
A.5	Opening the on-line help	249
A.6	The SAS Help and Documentation window	250

B.1	R Windows graphical user interface	262
B.2	R Mac OS X graphical user interface	263
B.3	Sample session in R	264
B.4	Documentation on the <code>mean()</code> function	267

List of Tables

1.1	Quantiles, probabilities, and pseudo-random number generation: distributions available in SAS and R	43
4.1	Generalized linear model distributions supported by SAS and R	132
C.1	Analyses undertaken using the HELP dataset	277
C.2	Annotated description of variables in the HELP dataset	279

