

Applied Statistics with R

2021-07-23

Contents

1	Introduction	11
1.1	About This Book	11
1.2	Conventions	12
1.3	Acknowledgements	12
1.4	License	14
2	Introduction to R	15
2.1	Getting Started	15
2.2	Basic Calculations	16
2.3	Getting Help	17
2.4	Installing Packages	18
3	Data and Programming	21
3.1	Data Types	21
3.2	Data Structures	21
3.2.1	Vectors	22
3.2.2	Vectorization	26
3.2.3	Logical Operators	27
3.2.4	More Vectorization	29
3.2.5	Matrices	31
3.2.6	Lists	41
3.2.7	Data Frames	43
3.3	Programming Basics	51

3.3.1	Control Flow	51
3.3.2	Functions	52
4	Summarizing Data	57
4.1	Summary Statistics	57
4.2	Plotting	58
4.2.1	Histograms	58
4.2.2	Barplots	60
4.2.3	Boxplots	62
4.2.4	Scatterplots	64
5	Probability and Statistics in R	67
5.1	Probability in R	67
5.1.1	Distributions	67
5.2	Hypothesis Tests in R	69
5.2.1	One Sample t-Test: Review	69
5.2.2	One Sample t-Test: Example	70
5.2.3	Two Sample t-Test: Review	73
5.2.4	Two Sample t-Test: Example	73
5.3	Simulation	76
5.3.1	Paired Differences	77
5.3.2	Distribution of a Sample Mean	80
6	R Resources	85
6.1	Beginner Tutorials and References	85
6.2	Intermediate References	85
6.3	Advanced References	86
6.4	Quick Comparisons to Other Languages	86
6.5	RStudio and RMarkdown Videos	86
6.6	RMarkdown Template	87

7	Simple Linear Regression	89
7.1	Modeling	89
7.1.1	Simple Linear Regression Model	94
7.2	Least Squares Approach	97
7.2.1	Making Predictions	99
7.2.2	Residuals	102
7.2.3	Variance Estimation	103
7.3	Decomposition of Variation	104
7.3.1	Coefficient of Determination	106
7.4	The <code>lm</code> Function	108
7.5	Maximum Likelihood Estimation (MLE) Approach	115
7.6	Simulating SLR	118
7.7	History	121
7.8	R Markdown	122
8	Inference for Simple Linear Regression	123
8.1	Gauss–Markov Theorem	126
8.2	Sampling Distributions	127
8.2.1	Simulating Sampling Distributions	128
8.3	Standard Errors	134
8.4	Confidence Intervals for Slope and Intercept	137
8.5	Hypothesis Tests	138
8.6	<code>cars</code> Example	139
8.6.1	Tests in R	139
8.6.2	Significance of Regression, t-Test	142
8.6.3	Confidence Intervals in R	143
8.7	Confidence Interval for Mean Response	145
8.8	Prediction Interval for New Observations	146
8.9	Confidence and Prediction Bands	147
8.10	Significance of Regression, F-Test	149
8.11	R Markdown	151

9 Multiple Linear Regression	153
9.1 Matrix Approach to Regression	157
9.2 Sampling Distribution	161
9.2.1 Single Parameter Tests	163
9.2.2 Confidence Intervals	165
9.2.3 Confidence Intervals for Mean Response	165
9.2.4 Prediction Intervals	169
9.3 Significance of Regression	170
9.4 Nested Models	174
9.5 Simulation	177
9.6 R Markdown	184
10 Model Building	185
10.1 Family, Form, and Fit	186
10.1.1 Fit	186
10.1.2 Form	187
10.1.3 Family	187
10.1.4 Assumed Model, Fitted Model	188
10.2 Explanation versus Prediction	189
10.2.1 Explanation	189
10.2.2 Prediction	191
10.3 Summary	194
10.4 R Markdown	194
11 Categorical Predictors and Interactions	195
11.1 Dummy Variables	196
11.2 Interactions	203
11.3 Factor Variables	212
11.3.1 Factors with More Than Two Levels	215
11.4 Parameterization	221
11.5 Building Larger Models	225
11.6 R Markdown	229

12 Analysis of Variance	231
12.1 Experiments	231
12.2 Two-Sample t-Test	232
12.3 One-Way ANOVA	235
12.3.1 Factor Variables	242
12.3.2 Some Simulation	243
12.3.3 Power	244
12.4 Post Hoc Testing	246
12.5 Two-Way ANOVA	249
12.6 R Markdown	259
 13 Model Diagnostics	 261
13.1 Model Assumptions	261
13.2 Checking Assumptions	263
13.2.1 Fitted versus Residuals Plot	264
13.2.2 Breusch-Pagan Test	270
13.2.3 Histograms	272
13.2.4 Q-Q Plots	273
13.2.5 Shapiro-Wilk Test	280
13.3 Unusual Observations	282
13.3.1 Leverage	284
13.3.2 Outliers	290
13.3.3 Influence	292
13.4 Data Analysis Examples	294
13.4.1 Good Diagnostics	294
13.4.2 Suspect Diagnostics	298
13.5 R Markdown	301

14 Transformations	303
14.1 Response Transformation	303
14.1.1 Variance Stabilizing Transformations	306
14.1.2 Box-Cox Transformations	311
14.2 Predictor Transformation	319
14.2.1 Polynomials	322
14.2.2 A Quadratic Model	345
14.2.3 Overfitting and Extrapolation	350
14.2.4 Comparing Polynomial Models	351
14.2.5 <code>poly()</code> Function and Orthogonal Polynomials	354
14.2.6 Inhibit Function	356
14.2.7 Data Example	357
14.3 R Markdown	363
15 Collinearity	365
15.1 Exact Collinearity	365
15.2 Collinearity	368
15.2.1 Variance Inflation Factor.	371
15.3 Simulation	377
15.4 R Markdown	382
16 Variable Selection and Model Building	383
16.1 Quality Criterion	383
16.1.1 Akaike Information Criterion	384
16.1.2 Bayesian Information Criterion	385
16.1.3 Adjusted R-Squared	386
16.1.4 Cross-Validated RMSE	386
16.2 Selection Procedures	390
16.2.1 Backward Search	391
16.2.2 Forward Search	397
16.2.3 Stepwise Search	400
16.2.4 Exhaustive Search	403

16.3 Higher Order Terms	408
16.4 Explanation versus Prediction	413
16.4.1 Explanation	413
16.4.2 Prediction	415
16.5 R Markdown	416
17 Logistic Regression	417
17.1 Generalized Linear Models	417
17.2 Binary Response	419
17.2.1 Fitting Logistic Regression	421
17.2.2 Fitting Issues	422
17.2.3 Simulation Examples	422
17.3 Working with Logistic Regression	429
17.3.1 Testing with GLMs	430
17.3.2 Wald Test	430
17.3.3 Likelihood-Ratio Test	431
17.3.4 SAheart Example	432
17.3.5 Confidence Intervals	435
17.3.6 Confidence Intervals for Mean Response	436
17.3.7 Formula Syntax	438
17.3.8 Deviance	440
17.4 Classification	441
17.4.1 spam Example	442
17.4.2 Evaluating Classifiers	445
17.5 R Markdown	452
18 Beyond	453
18.1 What's Next	453
18.2 RStudio	453
18.3 Tidy Data	453
18.4 Visualization	454
18.5 Web Applications	454

18.6 Experimental Design	454
18.7 Machine Learning	455
18.7.1 Deep Learning	455
18.8 Time Series	455
18.9 Bayesianism	456
18.10 High Performance Computing	456
18.11 Further R Resources	456
19 Appendix	457