Contents

Acknowledgments xvi
Preface xvii

1 Using and abusing data analytics in social science 1
1.1 Introduction 1
1.2 The promise of data analytics for social science 3
1.2.1 Data analytics in public affairs and public policy 3
1.2.2 Data analytics in the social sciences 3
1.2.3 Data analytics in the humanities 4
1.3 Research design issues in data analytics 4
1.3.1 Beware the true believer 4
1.3.2 Pseudo-objectivity in data analytics 4
1.3.3 The bias of scholarship based on algorithms using big data 5
1.3.4 The subjectivity of algorithms 8
1.3.5 Big data and big noise 9
1.3.6 Limitations of the leading data science dissemination models 9
1.4 Social and ethical issues in data analytics 10
1.4.1 Types of ethical issues in data analytics 10
1.4.2 Bias toward the privileged 11
1.4.3 Discrimination 12
1.4.4 Diversity and data analytics 13
1.4.5 Distortion of democratic processes 14
1.4.6 Undermining of professional ethics 14
1.4.7 Privacy, profiling, and surveillance issues 15
1.4.8 The transparency issue 18
1.5 Summary: Technology and power 19
Endnotes 21

2 Statistical analytics with R, Part 1 22
PART I: OVERVIEW OF STATISTICAL ANALYSIS WITH R 22
2.1 Introduction 22
2.2 Data and packages used in this chapter 22
2.2.1 Example data 22
2.2.2 R packages used 23
PART II: QUICK START ON STATISTICAL ANALYSIS WITH R 24
2.3 Descriptive statistics 24
2.4 Linear multiple regression 26
PART III: STATISTICAL ANALYSIS WITH R IN DETAIL 33
2.5 Hypothesis testing 33
2.5.1 One-sample test of means 34
2.5.2 Means test for two independent samples 35
2.5.3 Means test for two dependent samples 35
2.6 Crosstabulation, significance, and association 36
2.7 Loglinear analysis for categorical variables 38

2.8 Correlation, correlograms, and scatterplots 38
2.9 Factor analysis (exploratory) 43
2.10 Multidimensional scaling 44
2.11 Reliability analysis 44
2.11.1 Cronbach's alpha and Guttman's lower bounds 46
2.11.2 Guttman's lower bounds and Cronbach's alpha 46
2.11.3 Krippendorff's alpha and Cohen's kappa 48
2.12 Cluster analysis 49
2.12.1 Hierarchical cluster analysis 50
2.12.2 K-means clustering 50
2.12.3 Nearest neighbor analysis 59
2.13 Analysis of variance 60
2.13.1 Data and packages used 60
2.13.2 GLM univariate: ANOVA 61
2.13.3 GLM univariate: ANCOVA 66
2.13.4 GLM multivariate: MANOVA 67
2.13.5 GLM multivariate: MANCOVA 70
2.14 Logistic regression 73
2.14.1 ROC and AUC analysis 77
2.14.2 Confusion table and accuracy 77
2.15 Mediation and moderation 79
2.16 Chapter 2 command summary 89
Endnotes 89

3 Statistical analytics with R, Part 2 91

PART I: OVERVIEW OF STATISTICAL ANALYTICS WITH R 91
3.1 Introduction 91
3.2 Data and packages used in this chapter 91
3.2.1 Example data 91
3.2.2 R Packages used 92
PART II: QUICK START ON STATISTICAL ANALYSIS PART 2 92
3.3 Quick start: Linear regression as a generalized linear modeling (GZLM) 92
3.3.1 Background to GZLM 92
3.3.2 The linear model in `glm()` 92
3.3.3 GZLM output 93
3.3.4 Fitted value, residuals, and plots 94
3.3.5 Noncanonical custom links 97
3.3.6 Multiple comparison tests 98
3.3.7 Estimated marginal means (EMM) 98
3.4 Quick start: Testing if multilevel modeling is needed 99
PART III: STATISTICAL ANALYSIS, PART 2, IN DETAIL 101
3.5 Generalized linear models (GZLM) 101
3.5.1 Introduction 101
3.5.2 Setup for GZLM models in R 103
3.5.3 Binary logistic regression example 104
3.5.4 Gamma regression model 105
3.5.5 Poisson regression model 108
3.5.6 Negative binomial regression 113
3.6 Multilevel modeling (MLM) 115
3.6.1 Introduction 115
3.6.2 Setup and data 115

3.6.3 The random coefficients model 116
3.6.4 Likelihood ratio test 119
3.7 Panel data regression (PDR) 119
3.7.1 Introduction 119
3.7.2 Types of PDR model 120
3.7.3 The Hausman test 122
3.7.4 Setup and data 123
3.7.5 PDR with the plm package 124
3.7.6 PDR with the panelr package 133
3.8 Structural equation modeling (SEM) 134
3.9 Missing data analysis and data imputation 134
3.10 Chapter 3 command summary 134
Endnotes 134

4 Classification and regression trees in R 136
PART I: OVERVIEW OF CLASSIFICATION AND REGRESSION TREES WITH R 136
4.1 Introduction 137
4.2 Advantages of decision tree analysis 137
4.3 Limitations of decision tree analysis 138
4.4 Decision tree terminology 139
4.5 Steps in decision tree analysis 140
4.6 Decision tree algorithms 140
4.7 Random forests and ensemble methods 142
4.8 Software 143
4.8.1 R language 143
4.8.2 Stata 144
4.8.3 SAS 144
4.8.4 SPSS 144
4.8.5 Python language 144
4.9 Data and packages used in this chapter 144
4.9.1 Example data 144
4.9.2 R packages used 145
PART II: QUICK START - CLASSIFICATION AND REGRESSION TREES 145
4.10 Classification tree example: Survival on the Titanic 145
4.11 Regression tree example: Correlates of murder 149
PART III: CLASSIFICATION AND REGRESSION TREES, IN DETAIL 152
4.12 Overview 152
4.13 The `rpart()` program 153
4.13.1 Introduction 153
4.13.2 Training and validation datasets 155
4.13.3 Setup for `rpart()` trees 156
4.14 Classification trees with the rpart package 158
4.14.1 The basic rpart classification tree 158
4.14.2 Printing tree rules 160
4.14.3 Visualization with `prp()` and `draw.tree()` 161
4.14.4 Visualization with `fancyRpartPlot()` 163
4.14.5 Interpreting tree summaries 164
4.14.6 Listing nodes by country and countries by node 169
4.14.7 Node distribution plots 170
4.14.8 Saving predictions and residuals 171
4.14.9 Cross-validation and pruning 173

4.14.10 The confusion matrix and model performance metrics 176
4.14.11 The ROC curve and AUC 182
4.14.12 Lift plots 184
4.14.13 Gains plots 186
4.14.14 Precision vs. recall plot 186
4.15 Regression trees with the rpart package 189
4.15.1 Setup 189
4.15.2 Creating an rpart regression tree 189
4.15.3 Printing tree rules 192
4.15.4 Visualization with `prp()` and `fancyRpartPlot()` 192
4.15.5 Interpreting tree summaries 194
4.15.6 The CP table 197
4.15.7 Listing nodes by country and countries by node 198
4.15.8 Saving predictions and residuals 199
4.15.9 Plotting residuals 200
4.15.10 Cross-validation and pruning 201
4.15.11 R-squared for regression trees 202
4.15.12 MSE for regression trees 205
4.15.13 The confusion matrix 206
4.15.14 The ROC curve and AUC 206
4.15.15 Gains plots 206
4.15.16 Gains plot with OLS comparison 209
4.16 The tree package 212
4.17 The `ctree()` program for conditional decision trees 212
4.18 More decision trees programs for R 212
4.19 Chapter 4 command summary 213
Endnotes 213

5 Random forests 215
PART I: OVERVIEW OF RANDOM FORESTS IN R 215
5.1 Introduction 215
5.1.1 Social science examples of random forest models 215
5.1.2 Advantages of random forests 216
5.1.3 Limitations of random forests 217
5.1.4 Data and packages 217
PART II: QUICK START – RANDOM FORESTS 218
5.2 Classification forest example: Searching for the causes of happiness 218
5.3 Regression forest example: Why so much crime in my town? 221
PART III: RANDOM FORESTS, IN DETAIL 226
5.4 Classification forests with `randomForest()` 226
5.4.1 Setup 226
5.4.2 A basic classification model 227
5.4.3 Output components of `randomForest()` objects for classification models 230
5.4.4 Graphing a randomForest tree? 238
5.4.5 Comparing `randomForest()` and `rpart()` performance 239
5.4.6 Tuning the random forest model 241
5.4.7 MDS cluster analysis of the RF classification model 250
5.5 Regression forests with `randomForest()` 253
5.5.1 Introduction 253
5.5.2 Setup 254
5.5.3 A basic regression model 254
5.5.4 Output components for regression forest models 256

5.5.5 Graphing a randomForest tree? 260
5.5.6 MDS plots 260
5.5.7 Quartile plots 261
5.5.8 Comparing `randomForest()` and `rpart()` regression models 262
5.5.9 Tuning the `randomForest()` regression model 263
5.5.10 Outliers: Identifying and removing 268
5.6 The randomForestExplainer package 272
5.6.1 Setup for the randomForestExplainer package 272
5.6.2 Minimal depth plots 273
5.6.3 Multiway variable importance plots 274
5.6.4 Multiway ranking of variable importance 277
5.6.5 Comparing randomForest and OLS rankings of predictors 278
5.6.6 Which importance criteria? 280
5.6.7 Interaction analysis 281
5.6.8 The `explain_forest()` function 286
5.7 Summary 286
5.8 Conditional inference forests 287
5.9 MDS plots for random forests 287
5.10 More random forest programs for R 287
5.11 Command summary 289
Endnotes 289

6 Modeling and machine learning 291
PART I: OVERVIEW OF MODELING AND MACHINE LEARNING 291
6.1 Introduction 291
6.1.1 Social science examples of modeling and machine learning in R 292
6.1.2 Advantages of modeling and machine learning in R 294
6.1.3 Limitations of modeling and machine learning in R 294
6.1.4 Data, packages, and default directory 295
PART II: QUICK START – MODELING AND MACHINE LEARNING 297
6.2 Example 1: Bayesian modeling of county-level poverty 297
6.2.1 Introduction 297
6.2.2 Setup 297
6.2.3 Correlation plot 298
6.2.4 The Bayes generalized linear model 300
6.3 Example 2: Predicting diabetes among Pima Indians with mlr3 307
6.3.1 Introduction 307
6.3.2 Setup 307
6.3.3 How mlr3 works 307
6.3.4 The Pima Indian data 309
PART III: MODELING AND MACHINE LEARNING IN DETAIL 316
6.4 Illustrating modeling and machine learning with SVM in caret 316
6.4.1 How SVM works 317
6.4.2 SVM algorithms compared to logistic and OLS regression 317
6.4.3 SVM kernels, types, and parameters 318
6.4.4 Tuning SVM models 319
6.4.5 SVM and longitudinal data 319
6.5 SVM versus OLS regression 320
6.6 SVM with the caret package: Predicting world literacy rates 320
6.6.1 Setup 321
6.6.2 Constructing the SVM regression model with caret 322
6.6.3 Obtaining predicted values and residuals 323

6.6.4 Model performance metrics 323
6.6.5 Variable importance 324
6.6.6 Other output elements 324
6.6.7 SVM plots 325
6.7 Tuning SVM models 326
6.7.1 Tuning for the `train()` command from the caret package 327
6.7.2 Tuning for the `svm()` command from the e1071 package 328
6.7.3 Cross-validating SVM models 330
6.7.4 Using e1071 in caret rather than the default kern package 331
6.8 SVM classification models: Classifying U.S. Senators 333
6.8.1 The "senate" example and setup 333
6.8.2 SVM classification with alternative kernels: Senate example 333
6.8.3 Tuning the SVM binary classification model 338
6.9 Gradient boosting machines (GBM) 341
6.9.1 Introduction 341
6.9.2 Setup and example data 342
6.9.3 Metrics for comparing models 343
6.9.4 The caret control object 343
6.9.5 Training the GBM model under caret 344
6.10 Learning vector quantization (LVQ) 345
6.10.1 Introduction 345
6.10.2 Setup and example data 346
6.10.3 Metrics for comparing models 346
6.10.4 The caret control object 346
6.10.5 Training the LVQ model under caret 346
6.11 Comparing models 347
6.12 Variable importance 349
6.12.1 Leave-one-out modeling 349
6.12.2 Recursive feature elimination (RFE) with caret 350
6.12.3 Other approaches to variable importance 352
6.13 SVM classification for a multinomial outcome 352
6.14 Command summary 352
Endnotes 352

7 Neural network models and deep learning 355
PART I: OVERVIEW OF NEURAL NETWORK MODELS AND DEEP LEARNING 355
7.1 Overview 355
7.2 Data and packages 356
7.3 Social science examples 357
7.4 Pros and cons of neural networks 358
7.5 Artificial neural network (ANN) concepts 359
7.5.1 ANN terms 359
7.5.2 R software programs for ANN 362
7.5.3 Training methods for ANN 363
7.5.4 Algorithms in neuralnet 363
7.5.5 Algorithms in nnet 363
7.5.6 Tuning ANN models 364
PART II: QUICK START - MODELING AND MACHINE LEARNING 364
7.6 Example 1: Analyzing NYC airline delays 364
7.6.1 Introduction 364
7.6.2 General setup 364
7.6.3 Data preparation 364
7.6.4 Modeling NYC airline delays 365

7.7 Example 2: The classic iris classification example 370
7.7.1 Setup 370
7.7.2 Exploring separation with a violin plot 371
7.7.3 Normalizing the data 371
7.7.4 Training the model with nnet in caret 372
7.7.5 Obtain model predictions 374
7.7.6 Display the neural model 375
PART III: NEURAL NETWORK MODELS IN DETAIL 375
7.8 Analyzing Boston crime via the neuralnet package 375
7.8.1 Setup 376
7.8.2 The linear regression model for unscaled data 377
7.8.3 The neuralnet model for unscaled data 379
7.8.4 Scaling the data 379
7.8.5 The linear regression model for scaled data 379
7.8.6 The neuralnet model for scaled data 380
7.8.7 Neuralnet results for the training data 381
7.8.8 Model performance plots 382
7.8.9 Visualizing the neuralnet model 383
7.8.10 Variable importance for the neuralnet model 384
7.9 Analyzing Boston crime via neuralnet under the caret package 386
7.10 Analyzing Boston crime via nnet in caret 386
7.10.1 Setup 387
7.10.2 The nnet/caret model of Boston crime 388
7.10.3 Variable importance for the nnet/caret model 392
7.10.4 Further tuning the nnet model outside caret 393
7.11 A classification model of marital status using nnet 395
7.11.1 Setup 395
7.11.2 The nnet classification model of marital status 397
7.12 Neural network analysis using "mlr3keras" 400
7.13 Command summary 400
Endnotes 400

8 Network analysis 401

PART I: OVERVIEW OF NETWORK ANALYSIS WITH R 401
8.1 Introduction 401
8.2 Data and packages used in this chapter 401
8.3 Concepts in network analysis 403
8.4 Getting data into network format 404
PART II: QUICK START ON NETWORK ANALYSIS WITH R 405
8.5 Quick start exercise 1: The Medici family network 405
8.6 Quick start exercise 2: Marvel hero network communities 409
PART III: NETWORK ANALYSIS WITH R IN DETAIL 416
8.7 Interactive network analysis with visNetwork 416
8.7.1 Undirected networks: Research team management 417
8.7.2 Clustering by group: Research team grouped by gender 421
8.7.3 A larger network with navigation and circle layout 422
8.7.4 Visualizing classification and regression trees: National literacy 425
8.7.5 A directed network (asymmetrical relationships in a research team) 426
8.8 Network analysis with igraph 429
8.8.1 Term adjacency networks: Gubernatorial websites and the covid pandemic 429
8.8.2 Similarity/distance networks with igraph: Senate interest group ratings 436
8.8.3 Communities, modularity, and centrality 440
8.8.4 Similarity network analysis: All senators 447

8.9 Using intergraph for network conversions 453
8.10 Network-on-a-map with the diagram and maps packages 457
8.11 Network analysis with the statnet and network packages 462
8.11.1 Introduction 462
8.11.2 Visualization 467
8.11.3 Neighborhoods 470
8.11.4 Cluster analysis 472
8.12 Clique analysis with sna 473
8.12.1 A simplified clique analysis 473
8.12.2 A clique analysis of the DHHS formal network 475
8.12.3 K-core analysis of the DHHS formal network 481
8.13 Mapping international trade flow with statnet and Intergraph 481
8.14 Correlation networks with corrr 481
8.15 Network analysis with tidygraph 484
8.15.1 Introduction 484
8.15.2 A simple tidygraph example 484
8.15.3 Network conversions with tidygraph 490
8.15.4 Finding community clusters with tidygraph 491
8.16 Simulating networks 494
8.16.1 Agent-based network modeling with SchellingR 494
8.16.2 Agent-based network modeling with RSiena 499
8.16.3 Agent-based network modeling with NetLogoR 499
8.17 Summary 500
8.18 Command summary 501
Endnotes 501

9 Text analytics 503
PART I: OVERVIEW OF TEXT ANALYTICS WITH R 503
9.1 Overview 503
9.2 Data used in this chapter 503
9.3 Packages used in this chapter 504
9.4 What is a corpus? 505
9.5 Text files 505
9.5.1 Overview 505
9.5.2 Archived texts 505
9.5.3 Project Gutenberg archive 506
9.5.4 Comma-separated values (.csv) files 509
9.5.5 Text from Word .docx files with the textreadr package 509
9.5.6 Text from other formats with the readtext package 512
9.5.7 Text from raw text files 514
PART II: QUICK START ON TEXT ANALYTICS WITH R 516
9.6 Quick start exercise 1: Key word in context (kwic) indexing 516
9.7 Quick start exercise 2: Word frequencies and histograms 518
PART III: NETWORK ANALYSIS WITH R IN DETAIL 523
9.8 Web scraping 523
9.8.1 Overview 523
9.8.2 Web scraping: The "htm2txt" package 524
9.8.3 Web scraping: The "rvest" package 527
9.9 Social media scraping 531
9.9.1 Analysis of Twitter data: Trump and the *New York Times* 532
9.9.2 Social media scraping with twitter 536

9.10 Leading text formats in R 539
9.10.1 Overview 539
9.10.2 Formats related to the "tidytext" package 540
9.10.3 Formats related to the "tm" package 543
9.10.4 Formats related to the "quanteda" package 547
9.10.5 Common text file conversions 552
9.11 Tokenization 554
9.11.1 Overview 554
9.11.2 Word tokenization 554
9.12 Character encoding 557
9.13 Text cleaning and preparation 559
9.14 Analysis: Multigroup word frequency comparisons 559
9.14.1 Multigroup analysis in tidytext 559
9.14.2 Multigroup analysis with quanteda's `textstat_keyness()` command 563
9.14.3 Multigroup analysis with `textstat_frequency()` in quanteda and ggplot2 566
9.15 Analysis: Word clouds 567
9.16 Analysis: Comparison clouds 572
9.17 Analysis: Word maps and word correlations 574
9.17.1 Working with the tdm format 574
9.17.2 Working with the dtm format 575
9.17.3 Word frequencies and word correlations 576
9.17.4 Correlation plots of word and document associations 577
9.17.5 Plotting word stem correlations for word pairs 581
9.17.6 Word correlation maps 584
9.18 Analysis: Sentiment analysis 587
9.18.1 Overview 587
9.18.2 Example: Sentiment analysis of news articles 587
9.19 Analysis: Topic modeling 596
9.19.1 Overview 596
9.19.2 Topic analysis example 1: Modeling topic frequency over time 597
9.19.3 Topic analysis example 2: LDA analysis 603
9.20 Analysis: Lexical dispersion plots 610
9.21 Analysis: Bigrams and ngrams 611
9.22 Command summary 612
Endnotes 612

Appendix 1: Introduction to R and RStudio 613
Appendix 2: Data used in this book 658
References 668
Index 678