

Contents

Preface — V

Chapter 1

Data Science in Social Research — 1

- 1.1 Data Science — 1
- 1.2 Computational Methods in Social Research — 1
- 1.3 Web Survey Research — 3
- 1.4 Changing Survey Landscape — 4
- 1.5 Digital Trace Data — 5
- 1.6 Social Media Research — 7
- 1.7 Big Data — 8
- 1.8 Prediction and Explanation — 9
- 1.9 Outlook on the Chapters of this Textbook — 10
- 1.10 Example Data Sets Used in this Textbook — 14
- 1.11 Recommended Course Reading — 16

Part I: Description, Explanation, Prediction

Chapter 2

Descriptive Inference — 21

- 2.1 Descriptive and Statistical Inference — 21
- 2.2 Description by Population Parameters — 22
- 2.3 Selection Frame — 22
- 2.4 Probability Sampling — 24
- 2.5 Design-based Inference — 27
- 2.6 Model-Assisted, Design-Based Inference — 29
 - 2.6.1 Use Case: Supplied Nonresponse Weights — 29
 - 2.6.2 Use Case: Models of Response Propensity — 31
 - 2.6.3 Use Case: Quasi-Randomized Samples — 36
- 2.7 Missing Data — 37
 - 2.7.1 Patterns — 37
 - 2.7.2 Procedures — 38
 - 2.7.3 Multiple Imputation — 40
 - 2.7.4 Predictive Mean Matching — 42
 - 2.7.5 Example Data — 43
 - 2.7.6 Using the Library *mice* for a Predictive Mean Matching and Multiple Imputation — 43
- 2.8 Statistical Matching — 47

Chapter 3

Causal Inference — 50

- 3.1 Counterfactuals — 50
- 3.2 Comparing Like with Like — 51
- 3.3 Limits of Randomization and Matching — 53
- 3.4 Observational Studies — 55
 - 3.4.1 Statistical Conclusion Validity — 55
 - 3.4.2 Internal Validity — 56
 - 3.4.3 External Validity — 58
 - 3.4.4 Construct Validity — 59
 - 3.4.5 Analytically Controlling on Third Variables — 60
 - 3.4.6 Causal Inference — 68
 - 3.4.7 Unobserved Heterogeneity — 77
- 3.5 Replication — 81
 - 3.5.1 Conceptualization > Measurement — 82
 - 3.5.2 Measurement > Modeling — 83
 - 3.5.3 Modeling > Estimation — 84
 - 3.5.4 Estimation > Validation — 85

Part II: Using R

Chapter 4

Data Management in R — 89

- 4.1 R and R-Studio — 89
- 4.2 Importing Data From Excel and Stata — 90
- 4.3 The Elements of a Data Frame — 93
- 4.4 Univariate Frequency Distributions — 96
- 4.5 Vectors — 108
 - 4.5.1 Vector, Matrix, List, Data Frame — 108
 - 4.5.2 Objects and Functions — 108
 - 4.5.3 Using the c() and cbind() Functions — 109
 - 4.5.4 Vector Operations — 114
 - 4.5.5 Statistical Results as Numerical Vectors — 115
- 4.6 Recoding the Values of Variables with recode() and ifelse() — 120
- 4.7 Creating Subsets with subset() and sample() — 124
- 4.8 Statistical Calculations with Weighted Units — 126
- 4.9 Multiple Responses and Rankings — 128
- 4.10 Data Frame and Tibble — 132
- 4.11 Convert a Tibble From Wide to Long Format — 137
- 4.12 The Pipe Operator — 139
- 4.13 Tidy Text Format — 141

Chapter 5**R Graphics — 144**

- 5.1 R's Plot() Function — 144
- 5.2 Exploratory Graphs — 145
 - 5.2.1 Distributional Shape, Mean Values and Dispersion — 145
 - 5.2.2 Scatterplots — 148
- 5.3 Predictive Graphs — 159
- 5.4 Diagnostic Graphs — 164

Part III: Statistical Methods for Prediction, Classification, Text Analytics

Chapter 6**Linear Models — 175**

- 6.1 Simple Regression — 175
- 6.2 Polynomial Regression — 177
- 6.3 Multiple Regression — 178
- 6.4 Ordinary Least Squares — 178
- 6.5 Partial Regression Coefficients — 181
- 6.6 Weighting the Units of Analysis — 185
- 6.7 Statistical Interaction in the Standard Use Case — 187
- 6.8 Statistical Interaction in the Mixed-Scales Use Case — 189

Chapter 7**Goodness of Fit and Predictive Power of a Linear Model — 194**

- 7.1 Within-Sample Model Fit: How Well Does a Linear Model Fit the Data? — 194
- 7.2 Linear Regression in the Validation Set Design — 200
- 7.3 Validation Set and Temporal Validation Design in Structural Equation Modeling — 208
- 7.4 Cross-Validation — 218
- 7.5 Generalizing Across Samples, Methods, and Settings — 226

Chapter 8**Regression Diagnostics and Corrective Measures — 229**

- 8.1 Regression Diagnostics – Going Beyond R^2 — 229
- 8.2 Residual-vs.-Fitted Plot as Diagnostic Instrument — 231
- 8.3 Reducing Autocorrelation in Residuals by Linear Mixed Models — 236
- 8.4 Normally Distributed Residuals — 239
- 8.5 Specifying the Random Component by a Generalized Linear Model — 242

- 8.6 When the Variance of a Distribution Increases with its Mean: Assessing the Occurrence of Future Events Using the Gamma Distribution — **243**
- 8.7 When the Variance of a Distribution is a Function of its Mean: Using the Binomial Distribution — **253**
 - 8.7.1 Example Data — **254**
 - 8.7.2 Linear Probability, Logistic, and Probit Model — **255**
 - 8.7.3 Logistic Link Function — **262**
 - 8.7.4 Weighting the $Pr(y = 1|x)$ — **262**
 - 8.7.5 Inference With the Binomial Distribution — **263**

Chapter 9

Linear Mixed Models — 268

- 9.1 Fixed and Random Effects — **268**
- 9.2 Linear and Linear Mixed Model Compared — **270**
- 9.3 Comparative Model Evaluation — **274**
- 9.4 Variance Decomposition — **275**
- 9.5 Tables to the Models on Prejudice Formation and Authoritarian Personality — **277**

Chapter 10

Nonlinear Relations — 286

- 10.1 Linearizable Relations — **286**
- 10.2 Polynomial Regression and Regression Splines — **298**
- 10.3 Generalized Additive Models — **311**

Chapter 11

Classification — 324

- 11.1 Predicting Class Membership by Logistic Regression — **324**
- 11.2 Contingency Tables of Weighted Units of Analysis — **327**
 - 11.2.1 Example Data — **328**
 - 11.2.2 Weighting the Units of Analysis for *ftable()* Structures and LCA Input — **329**
- 11.3 Predicting Class Membership by WLS Models for Contingency Tables — **332**
- 11.4 Probabilistic Classification by Latent Class Analysis — **339**
 - 11.4.1 Latent Structure Analysis — **339**
 - 11.4.2 Basic Model — **340**
 - 11.4.3 Example Data — **341**
 - 11.4.4 Conditional Response Probabilities — **341**
 - 11.4.5 Latent Classes — **344**
 - 11.4.6 Model Fit — **345**

- 11.4.7 Posterior Probabilities and Predicted Class Memberships — 349
- 11.4.8 Expected Cell and Latent-Class Frequencies — 352
- 11.4.9 Entropy-based Classification Quality — 354

Chapter 12

Statistical Learning — 359

- 12.1 Statistical Learning — 359
- 12.2 Learning Functions from Data — 361
- 12.3 Modeling and Feature Engineering — 364
 - 12.3.1 Example Data — 365
 - 12.3.2 Using the *tidymodels* Metapackage — 370
 - 12.3.3 Test Step — 371
- 12.4 Regression and Classification Trees — 375
 - 12.4.1 Trees in the Feature Space — 376
 - 12.4.2 Example Data — 378
 - 12.4.3 Regression Trees — 379
 - 12.4.4 Classification Trees — 381
 - 12.4.5 Resampling Using the Bootstrap — 383
 - 12.4.6 Bagging, Random Forests, Boosting — 386
- 12.5 Artificial Neural Networks — 389

Chapter 13

Natural Language Processing — 395

- 13.1 Introduction — 395
- 13.2 Methods of Data Collection for Textual Data — 396
 - 13.2.1 Web Scraping — 397
 - 13.2.2 Application Programming Interfaces (APIs) — 398
- 13.3 Text Quality and Challenges — 401
- 13.4 Preprocessing of Textual Data — 402
- 13.5 Data Structures for Textual Data — 406
- 13.6 First Insights into Textual Data — 407
- 13.7 Topic Modeling — 411
 - 13.7.1 Introduction — 411
 - 13.7.2 Number of Topics and Interpretability — 412
 - 13.7.3 Common Topic Models — 413
 - 13.7.3.1 Latent Dirichlet Allocation (LDA) — 413
 - 13.7.3.2 Biterm Topic Model (BTM) — 414
 - 13.7.3.3 Structural Topic Model (STM) — 418
- 13.8 Sentiment Analysis — 428
 - 13.8.1 Introduction — 428
 - 13.8.2 Lexicon-based Sentiment Analysis — 428

References — 433

R Libraries Used — 443

List of Figures — 451

List of Tables — 453

About the Authors — 457

Subject Index — 459