Contents

Preface **xi**
What you should be able to do after reading and working through this book xi
Fun chapter titles xii
Additional material for teaching and learning xiii

Part 1: Fundamentals **1**

1 Overview **3**
1.1 The three challenges of statistics 3
1.2 Why learn regression? 4
1.3 Some examples of regression 5
1.4 Challenges in building, understanding, and interpreting regressions 9
1.5 Classical and Bayesian inference 13
1.6 Computing least squares and Bayesian regression 16
1.7 Bibliographic note 17
1.8 Exercises 17

2 Data and measurement **21**
2.1 Examining where data come from 21
2.2 Validity and reliability 23
2.3 All graphs are comparisons 25
2.4 Data and adjustment: trends in mortality rates 31
2.5 Bibliographic note 33
2.6 Exercises 34

3 Some basic methods in mathematics and probability **35**
3.1 Weighted averages 35
3.2 Vectors and matrices 36
3.3 Graphing a line 37
3.4 Exponential and power-law growth and decline; logarithmic and log-log relationships 38
3.5 Probability distributions 40
3.6 Probability modeling 45
3.7 Bibliographic note 47
3.8 Exercises 47

4 Statistical inference **49**
4.1 Sampling distributions and generative models 49
4.2 Estimates, standard errors, and confidence intervals 50
4.3 Bias and unmodeled uncertainty 55
4.4 Statistical significance, hypothesis testing, and statistical errors 57
4.5 Problems with the concept of statistical significance 60
4.6 Example of hypothesis testing: 55,000 residents need your help! 63
4.7 Moving beyond hypothesis testing 66

4.8 Bibliographic note 67
4.9 Exercises 67

5 Simulation 69
5.1 Simulation of discrete probability models 69
5.2 Simulation of continuous and mixed discrete/continuous models 71
5.3 Summarizing a set of simulations using median and median absolute deviation 73
5.4 Bootstrapping to simulate a sampling distribution 73
5.5 Fake-data simulation as a way of life 76
5.6 Bibliographic note 76
5.7 Exercises 76

Part 2: Linear regression 79

6 Background on regression modeling 81
6.1 Regression models 81
6.2 Fitting a simple regression to fake data 82
6.3 Interpret coefficients as comparisons, not effects 84
6.4 Historical origins of regression 85
6.5 The paradox of regression to the mean 87
6.6 Bibliographic note 90
6.7 Exercises 91

7 Linear regression with a single predictor 93
7.1 Example: predicting presidential vote share from the economy 93
7.2 Checking the model-fitting procedure using fake-data simulation 97
7.3 Formulating comparisons as regression models 99
7.4 Bibliographic note 101
7.5 Exercises 101

8 Fitting regression models 103
8.1 Least squares, maximum likelihood, and Bayesian inference 103
8.2 Influence of individual points in a fitted regression 107
8.3 Least squares slope as a weighted average of slopes of pairs 108
8.4 Comparing two fitting functions: lm and stan_glm 109
8.5 Bibliographic note 111
8.6 Exercises 111

9 Prediction and Bayesian inference 113
9.1 Propagating uncertainty in inference using posterior simulations 113
9.2 Prediction and uncertainty: predict, posterior_linpred, and posterior_predict 115
9.3 Prior information and Bayesian synthesis 119
9.4 Example of Bayesian inference: beauty and sex ratio 121
9.5 Uniform, weakly informative, and informative priors in regression 123
9.6 Bibliographic note 128
9.7 Exercises 128

10 Linear regression with multiple predictors 131
10.1 Adding predictors to a model 131
10.2 Interpreting regression coefficients 133
10.3 Interactions 134
10.4 Indicator variables 136
10.5 Formulating paired or blocked designs as a regression problem 139

10.6 Example: uncertainty in predicting congressional elections 140
10.7 Mathematical notation and statistical inference 144
10.8 Weighted regression 147
10.9 Fitting the same model to many datasets 148
10.10 Bibliographic note 149
10.11 Exercises 150

11 Assumptions, diagnostics, and model evaluation 153
11.1 Assumptions of regression analysis 153
11.2 Plotting the data and fitted model 156
11.3 Residual plots 161
11.4 Comparing data to replications from a fitted model 163
11.5 Example: predictive simulation to check the fit of a time-series model 166
11.6 Residual standard deviation σ and explained variance R^2 168
11.7 External validation: checking fitted model on new data 171
11.8 Cross validation 172
11.9 Bibliographic note 180
11.10 Exercises 180

12 Transformations and regression 183
12.1 Linear transformations 183
12.2 Centering and standardizing for models with interactions 185
12.3 Correlation and "regression to the mean" 187
12.4 Logarithmic transformations 189
12.5 Other transformations 195
12.6 Building and comparing regression models for prediction 199
12.7 Models for regression coefficients 206
12.8 Bibliographic note 210
12.9 Exercises 211

Part 3: Generalized linear models 215

13 Logistic regression 217
13.1 Logistic regression with a single predictor 217
13.2 Interpreting logistic regression coefficients and the divide-by-4 rule 220
13.3 Predictions and comparisons 222
13.4 Latent-data formulation 226
13.5 Maximum likelihood and Bayesian inference for logistic regression 228
13.6 Cross validation and log score for logistic regression 230
13.7 Building a logistic regression model: wells in Bangladesh 232
13.8 Bibliographic note 237
13.9 Exercises 237

14 Working with logistic regression 241
14.1 Graphing logistic regression and binary data 241
14.2 Logistic regression with interactions 242
14.3 Predictive simulation 247
14.4 Average predictive comparisons on the probability scale 249
14.5 Residuals for discrete-data regression 253
14.6 Identification and separation 256
14.7 Bibliographic note 259
14.8 Exercises 259

15 Other generalized linear models 263
15.1 Definition and notation 263
15.2 Poisson and negative binomial regression 264
15.3 Logistic-binomial model 270
15.4 Probit regression: normally distributed latent data 272
15.5 Ordered and unordered categorical regression 273
15.6 Robust regression using the *t* model 278
15.7 Constructive choice models 279
15.8 Going beyond generalized linear models 283
15.9 Bibliographic note 286
15.10 Exercises 286

Part 4: Before and after fitting a regression 289

16 Design and sample size decisions 291
16.1 The problem with statistical power 291
16.2 General principles of design, as illustrated by estimates of proportions 293
16.3 Sample size and design calculations for continuous outcomes 297
16.4 Interactions are harder to estimate than main effects 301
16.5 Design calculations after the data have been collected 304
16.6 Design analysis using fake-data simulation 306
16.7 Bibliographic note 310
16.8 Exercises 310

17 Poststratification and missing-data imputation 313
17.1 Poststratification: using regression to generalize to a new population 313
17.2 Fake-data simulation for regression and poststratification 320
17.3 Models for missingness 322
17.4 Simple approaches for handling missing data 324
17.5 Understanding multiple imputation 326
17.6 Nonignorable missing-data models 332
17.7 Bibliographic note 333
17.8 Exercises 333

Part 5: Causal inference 337

18 Causal inference and randomized experiments 339
18.1 Basics of causal inference 339
18.2 Average causal effects 342
18.3 Randomized experiments 345
18.4 Sampling distributions, randomization distributions, and bias in estimation 346
18.5 Using additional information in experimental design 347
18.6 Properties, assumptions, and limitations of randomized experiments 350
18.7 Bibliographic note 355
18.8 Exercises 356

19 Causal inference using regression on the treatment variable 363
19.1 Pre-treatment covariates, treatments, and potential outcomes 363
19.2 Example: the effect of showing children an educational television show 364
19.3 Including pre-treatment predictors 367
19.4 Varying treatment effects, interactions, and poststratification 370
19.5 Challenges of interpreting regression coefficients as treatment effects 373
19.6 Do not adjust for post-treatment variables 374

19.7 Intermediate outcomes and causal paths 376
19.8 Bibliographic note 379
19.9 Exercises 380

20 Observational studies with all confounders assumed to be measured 383
20.1 The challenge of causal inference 383
20.2 Using regression to estimate a causal effect from observational data 386
20.3 Assumption of ignorable treatment assignment in an observational study 388
20.4 Imbalance and lack of complete overlap 391
20.5 Example: evaluating a child care program 394
20.6 Subclassification and average treatment effects 397
20.7 Propensity score matching for the child care example 399
20.8 Restructuring to create balanced treatment and control groups 405
20.9 Additional considerations with observational studies 413
20.10 Bibliographic note 416
20.11 Exercises 417

21 Additional topics in causal inference 421
21.1 Estimating causal effects indirectly using instrumental variables 421
21.2 Instrumental variables in a regression framework 427
21.3 Regression discontinuity: known assignment mechanism but no overlap 432
21.4 Identification using variation within or between groups 440
21.5 Causes of effects and effects of causes 445
21.6 Bibliographic note 449
21.7 Exercises 450

Part 6: What comes next? 455

22 Advanced regression and multilevel models 457
22.1 Expressing the models so far in a common framework 457
22.2 Incomplete data 458
22.3 Correlated errors and multivariate models 459
22.4 Regularization for models with many predictors 459
22.5 Multilevel or hierarchical models 460
22.6 Nonlinear models, a demonstration using Stan 460
22.7 Nonparametric regression and machine learning 464
22.8 Computational efficiency 467
22.9 Bibliographic note 471
22.10 Exercises 471

Appendixes 473

A Computing in R 475
A.1 Downloading and installing R and Stan 475
A.2 Accessing data and code for the examples in the book 476
A.3 The basics 476
A.4 Reading, writing, and looking at data 481
A.5 Making graphs 482
A.6 Working with messy data 484
A.7 Some R programming 488
A.8 Working with rstanarm fit objects 490
A.9 Bibliographic note 492

B 10 quick tips to improve your regression modeling **493**
B.1 Think about variation and replication 493
B.2 Forget about statistical significance 493
B.3 Graph the relevant and not the irrelevant 493
B.4 Interpret regression coefficients as comparisons 494
B.5 Understand statistical methods using fake-data simulation 494
B.6 Fit many models 495
B.7 Set up a computational workflow 495
B.8 Use transformations 496
B.9 Do causal inference in a targeted way, not as a byproduct of a large regression 496
B.10 Learn methods through live examples 496

References **497**

Author Index **516**

Subject Index **523**