

# Contents

|                                                                              |            |
|------------------------------------------------------------------------------|------------|
| <b>Preface</b>                                                               | <b>xi</b>  |
| <b>1 Analyzing a binary response, part 1: introduction</b>                   | <b>1</b>   |
| 1.1 One binary variable . . . . .                                            | 1          |
| 1.1.1 Bernoulli and binomial probability distributions . . . . .             | 1          |
| 1.1.2 Inference for the probability of success . . . . .                     | 8          |
| 1.1.3 True confidence levels for confidence intervals . . . . .              | 17         |
| 1.2 Two binary variables . . . . .                                           | 23         |
| 1.2.1 Notation and model . . . . .                                           | 25         |
| 1.2.2 Confidence intervals for the difference of two probabilities . . . . . | 29         |
| 1.2.3 Test for the difference of two probabilities . . . . .                 | 34         |
| 1.2.4 Relative risks . . . . .                                               | 37         |
| 1.2.5 Odds ratios . . . . .                                                  | 40         |
| 1.2.6 Matched pairs data . . . . .                                           | 43         |
| 1.2.7 Larger contingency tables . . . . .                                    | 47         |
| 1.3 Exercises . . . . .                                                      | 48         |
| <b>2 Analyzing a binary response, part 2: regression models</b>              | <b>61</b>  |
| 2.1 Linear regression models . . . . .                                       | 61         |
| 2.2 Logistic regression models . . . . .                                     | 62         |
| 2.2.1 Parameter estimation . . . . .                                         | 64         |
| 2.2.2 Hypothesis tests for regression parameters . . . . .                   | 76         |
| 2.2.3 Odds ratios . . . . .                                                  | 82         |
| 2.2.4 Probability of success . . . . .                                       | 86         |
| 2.2.5 Interactions and transformations for explanatory variables . . . . .   | 94         |
| 2.2.6 Categorical explanatory variables . . . . .                            | 100        |
| 2.2.7 Convergence of parameter estimation . . . . .                          | 110        |
| 2.2.8 Monte Carlo simulation . . . . .                                       | 116        |
| 2.3 Generalized linear models . . . . .                                      | 121        |
| 2.4 Exercises . . . . .                                                      | 129        |
| <b>3 Analyzing a multicategory response</b>                                  | <b>141</b> |
| 3.1 Multinomial probability distribution . . . . .                           | 141        |
| 3.2 $I \times J$ contingency tables and inference procedures . . . . .       | 143        |
| 3.2.1 One multinomial distribution . . . . .                                 | 143        |
| 3.2.2 $I$ multinomial distributions . . . . .                                | 144        |
| 3.2.3 Test for independence . . . . .                                        | 146        |
| 3.3 Nominal response regression models . . . . .                             | 152        |
| 3.3.1 Odds ratios . . . . .                                                  | 160        |
| 3.3.2 Contingency tables . . . . .                                           | 163        |
| 3.4 Ordinal response regression models . . . . .                             | 170        |
| 3.4.1 Odds ratios . . . . .                                                  | 177        |
| 3.4.2 Contingency tables . . . . .                                           | 179        |

|          |                                                                       |            |
|----------|-----------------------------------------------------------------------|------------|
| 3.4.3    | Non-proportional odds model . . . . .                                 | 182        |
| 3.5      | Additional regression models . . . . .                                | 186        |
| 3.6      | Exercises . . . . .                                                   | 187        |
| <b>4</b> | <b>Analyzing a count response</b>                                     | <b>195</b> |
| 4.1      | Poisson model for count data . . . . .                                | 195        |
| 4.1.1    | Poisson distribution . . . . .                                        | 195        |
| 4.1.2    | Poisson likelihood and inference . . . . .                            | 198        |
| 4.2      | Poisson regression models for count responses . . . . .               | 201        |
| 4.2.1    | Model for mean: Log link . . . . .                                    | 203        |
| 4.2.2    | Parameter estimation and inference . . . . .                          | 204        |
| 4.2.3    | Categorical explanatory variables . . . . .                           | 211        |
| 4.2.4    | Poisson regression for contingency tables: loglinear models . . . . . | 218        |
| 4.2.5    | Large loglinear models . . . . .                                      | 222        |
| 4.2.6    | Ordinal categorical variables . . . . .                               | 231        |
| 4.3      | Poisson rate regression . . . . .                                     | 240        |
| 4.4      | Zero inflation . . . . .                                              | 244        |
| 4.5      | Exercises . . . . .                                                   | 253        |
| <b>5</b> | <b>Model selection and evaluation</b>                                 | <b>265</b> |
| 5.1      | Variable selection . . . . .                                          | 265        |
| 5.1.1    | Overview of variable selection . . . . .                              | 265        |
| 5.1.2    | Model comparison criteria . . . . .                                   | 267        |
| 5.1.3    | All-subsets regression . . . . .                                      | 268        |
| 5.1.4    | Stepwise variable selection . . . . .                                 | 272        |
| 5.1.5    | Modern variable selection methods . . . . .                           | 277        |
| 5.1.6    | Model averaging . . . . .                                             | 281        |
| 5.2      | Tools to assess model fit . . . . .                                   | 285        |
| 5.2.1    | Residuals . . . . .                                                   | 286        |
| 5.2.2    | Goodness of fit . . . . .                                             | 292        |
| 5.2.3    | Influence . . . . .                                                   | 296        |
| 5.2.4    | Diagnostics for multcategory response models . . . . .                | 301        |
| 5.3      | Overdispersion . . . . .                                              | 301        |
| 5.3.1    | Causes and implications . . . . .                                     | 302        |
| 5.3.2    | Detection . . . . .                                                   | 305        |
| 5.3.3    | Solutions . . . . .                                                   | 306        |
| 5.4      | Examples . . . . .                                                    | 318        |
| 5.4.1    | Logistic regression - placekicking data set . . . . .                 | 318        |
| 5.4.2    | Poisson regression - alcohol consumption data set . . . . .           | 329        |
| 5.5      | Exercises . . . . .                                                   | 345        |
| <b>6</b> | <b>Additional topics</b>                                              | <b>355</b> |
| 6.1      | Binary responses and testing error . . . . .                          | 355        |
| 6.1.1    | Estimating the probability of success . . . . .                       | 355        |
| 6.1.2    | Binary regression models . . . . .                                    | 358        |
| 6.1.3    | Other methods . . . . .                                               | 361        |
| 6.2      | Exact inference . . . . .                                             | 362        |
| 6.2.1    | Fisher's exact test for independence . . . . .                        | 362        |
| 6.2.2    | Permutation test for independence . . . . .                           | 367        |
| 6.2.3    | Exact logistic regression . . . . .                                   | 372        |
| 6.2.4    | Additional exact inference procedures . . . . .                       | 380        |

|          |                                                           |            |
|----------|-----------------------------------------------------------|------------|
| 6.3      | Categorical data analysis in complex survey designs       | 380        |
| 6.3.1    | The survey sampling paradigm                              | 381        |
| 6.3.2    | Overview of analysis approaches                           | 382        |
| 6.3.3    | Weighted cell counts                                      | 386        |
| 6.3.4    | Inference on population proportions                       | 387        |
| 6.3.5    | Contingency tables and loglinear models                   | 389        |
| 6.3.6    | Logistic regression                                       | 400        |
| 6.4      | “Choose all that apply” data                              | 404        |
| 6.4.1    | Item response table                                       | 405        |
| 6.4.2    | Testing for marginal independence                         | 405        |
| 6.4.3    | Regression modeling                                       | 412        |
| 6.5      | Mixed models and estimating equations for correlated data | 419        |
| 6.5.1    | Random effects                                            | 420        |
| 6.5.2    | Mixed-effects models                                      | 422        |
| 6.5.3    | Model fitting                                             | 425        |
| 6.5.4    | Inference                                                 | 432        |
| 6.5.5    | Marginal modeling using generalized estimating equations  | 438        |
| 6.6      | Bayesian methods for categorical data                     | 443        |
| 6.6.1    | Estimating a probability of success                       | 444        |
| 6.6.2    | Regression models                                         | 449        |
| 6.6.3    | Alternative computational tools                           | 459        |
| 6.7      | Exercises                                                 | 459        |
| <b>A</b> | <b>An introduction to R</b>                               | <b>473</b> |
| A.1      | Basics                                                    | 473        |
| A.2      | Functions                                                 | 475        |
| A.3      | Help                                                      | 476        |
| A.4      | Using functions on vectors                                | 477        |
| A.5      | Packages                                                  | 478        |
| A.6      | Program editors                                           | 479        |
| A.6.1    | R editor                                                  | 479        |
| A.6.2    | RStudio                                                   | 479        |
| A.6.3    | Tinn-R                                                    | 480        |
| A.6.4    | Other editors                                             | 482        |
| A.7      | Regression example                                        | 482        |
| A.7.1    | Background                                                | 482        |
| A.7.2    | Data summary                                              | 483        |
| A.7.3    | Regression modeling                                       | 485        |
| A.7.4    | Additional items                                          | 491        |
| <b>B</b> | <b>Likelihood methods</b>                                 | <b>495</b> |
| B.1      | Introduction                                              | 495        |
| B.1.1    | Model and parameters                                      | 495        |
| B.1.2    | The role of likelihoods                                   | 496        |
| B.2      | Likelihood                                                | 496        |
| B.2.1    | Definition                                                | 496        |
| B.2.2    | Examples                                                  | 497        |
| B.3      | Maximum likelihood estimates                              | 498        |
| B.3.1    | Mathematical maximization of the log-likelihood function  | 500        |
| B.3.2    | Computational maximization of the log-likelihood function | 501        |
| B.3.3    | Large-sample properties of the MLE                        | 502        |

|       |                                         |            |
|-------|-----------------------------------------|------------|
| B.3.4 | Variance of the MLE                     | 502        |
| B.4   | Functions of parameters                 | 504        |
| B.4.1 | Invariance property of MLEs             | 504        |
| B.4.2 | Delta method for variances of functions | 505        |
| B.5   | Inference with MLEs                     | 506        |
| B.5.1 | Tests for parameters                    | 506        |
| B.5.2 | Confidence intervals for parameters     | 510        |
| B.5.3 | Tests for models                        | 511        |
|       | <b>Bibliography</b>                     | <b>513</b> |
|       | <b>Index</b>                            | <b>525</b> |