Table of Contents

Introduction — IX

1 Some fundamentals of empirical research — 1
1.1 Introduction — 1
1.2 On the relevance of quantitative methods in linguistics — 4
1.3 The design and the logic of quantitative studies — 9
1.3.1 Scouting — 10
1.3.2 Hypotheses and operationalization — 12
1.3.2.1 Scientific hypotheses in text form — 12
1.3.2.2 Operationalizing your variables — 16
1.3.2.3 Scientific hypotheses in statistical/mathematical form — 20
1.4 Data collection and storage — 24
1.5 The decision — 28
1.5.1 One-tailed *p*-values from discrete probability distributions — 31
1.5.2 Two-tailed *p*-values from discrete probability distributions — 36
1.5.3 Significance and effect size — 42
1.6 Data collection and storage — 44

2 Fundamentals of R — 51
2.1 Introduction and installation — 51
2.2 Functions and arguments — 54
2.3 Vectors — 57
2.3.1 Generating vectors — 57
2.3.2 Loading and saving vectors — 63
2.3.3 Working with vectors — 65
2.4 Factors — 74
2.4.1 Generating factors — 74
2.4.2 Loading and saving factors — 75
2.4.3 Working with factors — 76
2.5 Data frames — 77
2.5.1 Generating data frames — 78
2.5.2 Loading and saving data frames — 80
2.5.3 Working with data frames — 82
2.6 Lists — 87

3 Descriptive statistics — 91

3.1 Univariate descriptive statistics — 91
3.1.1 Categorical variables — 93
3.1.1.1 Central tendency: the mode — 94
3.1.1.2 Dispersion: normalized entropy — 95
3.1.1.3 Visualization — 96
3.1.2 Ordinal variables — 99
3.1.2.1 Central tendency: the median — 100
3.1.2.2 Dispersion: quantiles etc. — 101
3.1.2.3 Visualization — 102
3.1.3 Numeric variables — 103
3.1.3.1 Central tendency: arithmetic mean — 104
3.1.3.2 Dispersion: standard deviation etc. — 104
3.1.3.3 Visualization — 105
3.1.3.4 Two frequent transformations — 113
3.1.4 Standard errors and confidence intervals — 115
3.1.4.1 Standard errors for percentages — 116
3.1.4.2 Standard errors for means — 117
3.1.4.3 Confidence intervals — 118
3.2 Bivariate descriptive statistics — 121
3.2.1 Categorical/ordinal as a function of categorical/ordinal variables — 121
3.2.2 Categorical/ordinal variables as a function of numeric variables — 126
3.2.3 Numeric variables as a function of categorical/ordinal variables — 128
3.2.4 Numeric variables as a function of numeric variables — 129
3.3 Polemic excursus 1: on 'correlation' — 141
3.4 Polemic excursus 2: on visualization — 145
3.5 (Non-polemic) Excursus on programming — 148
3.5.1 Conditional expressions — 148
3.5.2 On looping — 152
3.5.3 On not looping: the apply family — 156
3.5.4 Function writing — 158
3.5.4.1 Anonymous functions — 159
3.5.4.2 Named functions — 160

4 Monofactorial tests — 164
4.1 Distributions and frequencies — 169
4.1.1 Goodness-of-fit — 169
4.1.1.1 One categorical/ordinal response — 169

4.1.1.2 One numeric response — 175
4.1.2 Tests for differences/independence — 178
4.1.2.1 One categorical response and one categorical predictor (indep.samples) — 178
4.1.2.2 One ordinal/numeric response and one categorical predictor (indep. samples) — 185
4.2 Dispersion — 189
4.2.1 Goodness-of-fit test for one numeric response — 190
4.2.2 Test for independence for one numeric response and one categorical predictor — 192
4.2.2.1 A small excursus: simulation — 199
4.3 Central tendencies — 200
4.3.1 Goodness-of-fit tests — 200
4.3.1.1 One ordinal response — 200
4.3.1.2 One numeric response — 205
4.3.2 Tests for differences/independence — 209
4.3.2.1 One ordinal response and one categorical predictor (indep. samples) — 209
4.3.2.2 One ordinal response and one categorical predictor (dep. samples) — 213
4.3.2.3 One numeric response and one categorical predictor (indep. samples) — 217
4.3.2.4 One numeric response and one categorical predictor (dep. samples) — 222
4.4 Correlation and simple linear regression — 227
4.4.1 Ordinal variables — 227
4.4.2 Numeric variables — 230
4.4.3 Correlation and causality — 233

5 Fixed-effects regression modeling — 235
5.1 A bit on 'multifactoriality' — 235
5.2 Linear regression — 240
5.2.1 A linear model with a numeric predictor — 241
5.2.1.1 Numerical exploration — 244
5.2.1.2 Graphical model exploration — 250
5.2.1.3 Excursus: curvature and anova — 251
5.2.1.4 Excursus: model frames and model matrices — 255
5.2.1.5 Excursus: the 95%-CI of the slope — 256
5.2.2 A linear model with a binary predictor — 257
5.2.2.1 Numerical exploration — 258

5.2.2.2 Graphical model exploration — 261
5.2.2.3 Excursus: coefficients as instructions — 262
5.2.3 A linear model with a categorical predictor — 263
5.2.3.1 Numerical exploration — 264
5.2.3.2 Graphical model exploration — 269
5.2.3.3 Excursus: conflation, model comparison, and contrasts — 270
5.2.4 Towards multifactorial modeling — 279
5.2.4.1 Simpson's paradox — 279
5.2.4.2 Interactions — 280
5.2.5 A linear model with two categorical predictors — 289
5.2.5.1 Numerical exploration — 290
5.2.5.2 Graphical model exploration — 294
5.2.5.3 Excursus: collinearity and *VIF*s — 296
5.2.6 A linear model with a categorical and a numeric predictor — 299
5.2.6.1 Numerical exploration — 299
5.2.6.2 Graphical model exploration — 302
5.2.6.3 Excursus: *post-hoc* comparisons and predictions from effects — 302
5.2.7 A linear model with two numeric predictors — 308
5.2.7.1 Numerical exploration — 309
5.2.7.2 Graphical model exploration — 311
5.2.7.3 Excursus: where are most of the values? — 313
5.2.8 Interactions (yes, again) — 314
5.3 Binary logistic regression — 319
5.3.1 A binary logistic regression with a binary predictor — 319
5.3.1.1 Numerical exploration — 319
5.3.1.2 Graphical model exploration — 327
5.3.2 A binary logistic regression with a categorical predictor — 328
5.3.2.1 Numerical exploration — 330
5.3.2.2 Graphical model exploration — 331
5.3.3 A binary logistic regression with a numeric predictor — 332
5.3.3.1 Numerical exploration — 333
5.3.3.2 Graphical model exploration — 335
5.3.3.3 Excursus: on cut-off points — 336
5.3.4 A binary logistic regression with two categorical predictors — 338
5.3.4.1 Numerical exploration — 339
5.3.4.2 Graphical model exploration — 341
5.3.5 Two more effects plots for you to recreate — 341
5.4 Other regression models — 343
5.4.1 Multinomial regression — 344

5.4.1.1 A multinomial regression with a numeric predictor — 345
5.4.1.2 A multinomial regression with a categorical predictor — 350
5.4.1.3 Multinomial and binary logistic regression — 352
5.4.2 Ordinal logistic regression — 353
5.4.2.1 An ordinal regression with a numeric predictor — 354
5.4.2.2 An ordinal regression with a categorical predictor — 358
5.5 Model formulation (and model selection) — 361
5.6 Model assumptions/diagnostics — 370
5.6.1 Amount of data — 371
5.6.2 Residuals — 372
5.6.3 Influential data points — 375
5.6.4 Excursus: autocorrelation/time & overdispersion — 379
5.7 Model validation (and classification vs. prediction) — 384
5.8 A thought experiment — 387

6 Mixed-effects regression modeling — 393
6.1 A very basic introduction — 393
6.1.1 Varying intercepts only — 397
6.1.2 Varying slopes only — 401
6.1.3 Varying intercepts and slopes — 403
6.1.3.1 Varying intercepts and slopes (correlated) — 404
6.1.3.2 Varying intercepts and slopes (uncorrelated) — 406
6.2 Some general MEM considerations — 408
6.3 Linear MEM case study — 414
6.3.1 Preparation and exploration — 414
6.3.2 Model fitting/selection — 418
6.3.3 Quick excursus on update — 423
6.3.4 Model diagnostics — 424
6.3.5 Model fitting/selection, part 2 — 425
6.3.6 A brief interlude — 428
6.3.7 Model diagnostics, part 2 — 430
6.3.8 Model interpretation — 432
6.3.9 A bit on MEM predictions — 437
6.4 Generalized linear MEM case study — 439
6.4.1 Preparation and exploration — 440
6.4.2 Model fitting/selection — 442
6.4.3 Model diagnostics — 445
6.4.4 Model interpretation — 445
6.5 On convergence and final recommendations — 450

7 Tree-based approaches — 453
7.1 Trees — 454
7.1.1 Classification and regression trees — 454
7.1.2 Conditional inference trees — 460
7.2 Ensembles of trees: forests — 463
7.2.1 Forests of classification and regression trees — 465
7.2.2 Forests of conditional inference trees — 469
7.3 Discussion — 471

References — 486

About the Author — 496