

AN INTRODUCTION TO CATEGORICAL DATA ANALYSIS

SECOND EDITION



ALAN AGRESTI

Contents

Preface to the Second Edition

xv

1. Introduction

1

- 1.1 Categorical Response Data, 1
 - 1.1.1 Response/Explanatory Variable Distinction, 2
 - 1.1.2 Nominal/Ordinal Scale Distinction, 2
 - 1.1.3 Organization of this Book, 3
- 1.2 Probability Distributions for Categorical Data, 3
 - 1.2.1 Binomial Distribution, 4
 - 1.2.2 Multinomial Distribution, 5
- 1.3 Statistical Inference for a Proportion, 6
 - 1.3.1 Likelihood Function and Maximum Likelihood Estimation, 6
 - 1.3.2 Significance Test About a Binomial Proportion, 8
 - 1.3.3 Example: Survey Results on Legalizing Abortion, 8
 - 1.3.4 Confidence Intervals for a Binomial Proportion, 9
- 1.4 More on Statistical Inference for Discrete Data, 11
 - 1.4.1 Wald, Likelihood-Ratio, and Score Inference, 11
 - 1.4.2 Wald, Score, and Likelihood-Ratio Inference for Binomial Parameter, 12
 - 1.4.3 Small-Sample Binomial Inference, 13
 - 1.4.4 Small-Sample Discrete Inference is Conservative, 14
 - 1.4.5 Inference Based on the Mid P -value, 15
 - 1.4.6 Summary, 16
- Problems, 16

2. Contingency Tables

21

- 2.1 Probability Structure for Contingency Tables, 21
 - 2.1.1 Joint, Marginal, and Conditional Probabilities, 22
 - 2.1.2 Example: Belief in Afterlife, 22

2.1.3	Sensitivity and Specificity in Diagnostic Tests, 23
2.1.4	Independence, 24
2.1.5	Binomial and Multinomial Sampling, 25
2.2	Comparing Proportions in Two-by-Two Tables, 25
2.2.1	Difference of Proportions, 26
2.2.2	Example: Aspirin and Heart Attacks, 26
2.2.3	Relative Risk, 27
2.3	The Odds Ratio, 28
2.3.1	Properties of the Odds Ratio, 29
2.3.2	Example: Odds Ratio for Aspirin Use and Heart Attacks, 30
2.3.3	Inference for Odds Ratios and Log Odds Ratios, 30
2.3.4	Relationship Between Odds Ratio and Relative Risk, 32
2.3.5	The Odds Ratio Applies in Case–Control Studies, 32
2.3.6	Types of Observational Studies, 34
2.4	Chi-Squared Tests of Independence, 34
2.4.1	Pearson Statistic and the Chi-Squared Distribution, 35
2.4.2	Likelihood-Ratio Statistic, 36
2.4.3	Tests of Independence, 36
2.4.4	Example: Gender Gap in Political Affiliation, 37
2.4.5	Residuals for Cells in a Contingency Table, 38
2.4.6	Partitioning Chi-Squared, 39
2.4.7	Comments About Chi-Squared Tests, 40
2.5	Testing Independence for Ordinal Data, 41
2.5.1	Linear Trend Alternative to Independence, 41
2.5.2	Example: Alcohol Use and Infant Malformation, 42
2.5.3	Extra Power with Ordinal Tests, 43
2.5.4	Choice of Scores, 43
2.5.5	Trend Tests for $I \times 2$ and $2 \times J$ Tables, 44
2.5.6	Nominal–Ordinal Tables, 45
2.6	Exact Inference for Small Samples, 45
2.6.1	Fisher’s Exact Test for 2×2 Tables, 45
2.6.2	Example: Fisher’s Tea Taster, 46
2.6.3	P -values and Conservatism for Actual P (Type I Error), 47
2.6.4	Small-Sample Confidence Interval for Odds Ratio, 48
2.7	Association in Three-Way Tables, 49
2.7.1	Partial Tables, 49
2.7.2	Conditional Versus Marginal Associations: Death Penalty Example, 49
2.7.3	Simpson’s Paradox, 51
2.7.4	Conditional and Marginal Odds Ratios, 52
2.7.5	Conditional Independence Versus Marginal Independence, 53
2.7.6	Homogeneous Association, 54
	Problems, 55

3. Generalized Linear Models**65**

- 3.1 Components of a Generalized Linear Model, 66
 - 3.1.1 Random Component, 66
 - 3.1.2 Systematic Component, 66
 - 3.1.3 Link Function, 66
 - 3.1.4 Normal GLM, 67
 - 3.2 Generalized Linear Models for Binary Data, 68
 - 3.2.1 Linear Probability Model, 68
 - 3.2.2 Example: Snoring and Heart Disease, 69
 - 3.2.3 Logistic Regression Model, 70
 - 3.2.4 Probit Regression Model, 72
 - 3.2.5 Binary Regression and Cumulative Distribution Functions, 72
 - 3.3 Generalized Linear Models for Count Data, 74
 - 3.3.1 Poisson Regression, 75
 - 3.3.2 Example: Female Horseshoe Crabs and their Satellites, 75
 - 3.3.3 Overdispersion: Greater Variability than Expected, 80
 - 3.3.4 Negative Binomial Regression, 81
 - 3.3.5 Count Regression for Rate Data, 82
 - 3.3.6 Example: British Train Accidents over Time, 83
 - 3.4 Statistical Inference and Model Checking, 84
 - 3.4.1 Inference about Model Parameters, 84
 - 3.4.2 Example: Snoring and Heart Disease Revisited, 85
 - 3.4.3 The Deviance, 85
 - 3.4.4 *Model Comparison Using the Deviance*, 86
 - 3.4.5 Residuals Comparing Observations to the Model Fit, 87
 - 3.5 Fitting Generalized Linear Models, 88
 - 3.5.1 The Newton–Raphson Algorithm Fits GLMs, 88
 - 3.5.2 Wald, Likelihood-Ratio, and Score Inference Use the Likelihood Function, 89
 - 3.5.3 Advantages of GLMs, 90
- Problems, 90

4. Logistic Regression**99**

- 4.1 Interpreting the Logistic Regression Model, 99
 - 4.1.1 Linear Approximation Interpretations, 100
 - 4.1.2 Horseshoe Crabs: Viewing and Smoothing a Binary Outcome, 101
 - 4.1.3 Horseshoe Crabs: Interpreting the Logistic Regression Fit, 101
 - 4.1.4 Odds Ratio Interpretation, 104

- 4.1.5 Logistic Regression with Retrospective Studies, 105
- 4.1.6 Normally Distributed X Implies Logistic Regression for Y , 105
- 4.2 Inference for Logistic Regression, 106
 - 4.2.1 Binary Data can be Grouped or Ungrouped, 106
 - 4.2.2 Confidence Intervals for Effects, 106
 - 4.2.3 Significance Testing, 107
 - 4.2.4 Confidence Intervals for Probabilities, 108
 - 4.2.5 Why Use a Model to Estimate Probabilities?, 108
 - 4.2.6 Confidence Intervals for Probabilities: Details, 108
 - 4.2.7 Standard Errors of Model Parameter Estimates, 109
- 4.3 Logistic Regression with Categorical Predictors, 110
 - 4.3.1 Indicator Variables Represent Categories of Predictors, 110
 - 4.3.2 Example: AZT Use and AIDS, 111
 - 4.3.3 ANOVA-Type Model Representation of Factors, 113
 - 4.3.4 The Cochran–Mantel–Haenszel Test for $2 \times 2 \times K$ Contingency Tables, 114
 - 4.3.5 Testing the Homogeneity of Odds Ratios, 115
- 4.4 Multiple Logistic Regression, 115
 - 4.4.1 Example: Horseshoe Crabs with Color and Width Predictors, 116
 - 4.4.2 Model Comparison to Check Whether a Term is Needed, 118
 - 4.4.3 Quantitative Treatment of Ordinal Predictor, 118
 - 4.4.4 Allowing Interaction, 119
- 4.5 Summarizing Effects in Logistic Regression, 120
 - 4.5.1 Probability-Based Interpretations, 120
 - 4.5.2 Standardized Interpretations, 121
- Problems, 121

5. Building and Applying Logistic Regression Models

137

- 5.1 Strategies in Model Selection, 137
 - 5.1.1 How Many Predictors Can You Use?, 138
 - 5.1.2 Example: Horseshoe Crabs Revisited, 138
 - 5.1.3 Stepwise Variable Selection Algorithms, 139
 - 5.1.4 Example: Backward Elimination for Horseshoe Crabs, 140
 - 5.1.5 AIC, Model Selection, and the “Correct” Model, 141
 - 5.1.6 Summarizing Predictive Power: Classification Tables, 142
 - 5.1.7 Summarizing Predictive Power: ROC Curves, 143
 - 5.1.8 Summarizing Predictive Power: A Correlation, 144
- 5.2 Model Checking, 144
 - 5.2.1 Likelihood-Ratio Model Comparison Tests, 144
 - 5.2.2 Goodness of Fit and the Deviance, 145

- 5.2.3 Checking Fit: Grouped Data, Ungrouped Data, and Continuous Predictors, 146
- 5.2.4 Residuals for Logit Models, 147
- 5.2.5 Example: Graduate Admissions at University of Florida, 149
- 5.2.6 Influence Diagnostics for Logistic Regression, 150
- 5.2.7 Example: Heart Disease and Blood Pressure, 151
- 5.3 Effects of Sparse Data, 152
 - 5.3.1 Infinite Effect Estimate: Quantitative Predictor, 152
 - 5.3.2 Infinite Effect Estimate: Categorical Predictors, 153
 - 5.3.3 Example: Clinical Trial with Sparse Data, 154
 - 5.3.4 Effect of Small Samples on X^2 and G^2 Tests, 156
- 5.4 Conditional Logistic Regression and Exact Inference, 157
 - 5.4.1 Conditional Maximum Likelihood Inference, 157
 - 5.4.2 Small-Sample Tests for Contingency Tables, 158
 - 5.4.3 Example: Promotion Discrimination, 159
 - 5.4.4 Small-Sample Confidence Intervals for Logistic Parameters and Odds Ratios, 159
 - 5.4.5 Limitations of Small-Sample Exact Methods, 160
- 5.5 Sample Size and Power for Logistic Regression, 160
 - 5.5.1 Sample Size for Comparing Two Proportions, 161
 - 5.5.2 Sample Size in Logistic Regression, 161
 - 5.5.3 Sample Size in Multiple Logistic Regression, 162
- Problems, 163

6. Multicategory Logit Models

173

- 6.1 Logit Models for Nominal Responses, 173
 - 6.1.1 Baseline-Category Logits, 173
 - 6.1.2 Example: Alligator Food Choice, 174
 - 6.1.3 Estimating Response Probabilities, 176
 - 6.1.4 Example: Belief in Afterlife, 178
 - 6.1.5 Discrete Choice Models, 179
- 6.2 Cumulative Logit Models for Ordinal Responses, 180
 - 6.2.1 Cumulative Logit Models with Proportional Odds Property, 180
 - 6.2.2 Example: Political Ideology and Party Affiliation, 182
 - 6.2.3 Inference about Model Parameters, 184
 - 6.2.4 Checking Model Fit, 184
 - 6.2.5 Example: Modeling Mental Health, 185
 - 6.2.6 Interpretations Comparing Cumulative Probabilities, 187
 - 6.2.7 Latent Variable Motivation, 187
 - 6.2.8 Invariance to Choice of Response Categories, 189
- 6.3 Paired-Category Ordinal Logits, 189

- 6.3.1 Adjacent-Categories Logits, 190
- 6.3.2 Example: Political Ideology Revisited, 190
- 6.3.3 Continuation-Ratio Logits, 191
- 6.3.4 Example: A Developmental Toxicity Study, 191
- 6.3.5 Overdispersion in Clustered Data, 192
- 6.4 Tests of Conditional Independence, 193
 - 6.4.1 Example: Job Satisfaction and Income, 193
 - 6.4.2 Generalized Cochran–Mantel–Haenszel Tests, 194
 - 6.4.3 Detecting Nominal–Ordinal Conditional Association, 195
 - 6.4.4 Detecting Nominal–Nominal Conditional Association, 196
- Problems, 196

7. Loglinear Models for Contingency Tables 204

- 7.1 Loglinear Models for Two-Way and Three-Way Tables, 204
 - 7.1.1 Loglinear Model of Independence for Two-Way Table, 205
 - 7.1.2 Interpretation of Parameters in Independence Model, 205
 - 7.1.3 Saturated Model for Two-Way Tables, 206
 - 7.1.4 Loglinear Models for Three-Way Tables, 208
 - 7.1.5 Two-Factor Parameters Describe Conditional Associations, 209
 - 7.1.6 Example: Alcohol, Cigarette, and Marijuana Use, 209
- 7.2 Inference for Loglinear Models, 212
 - 7.2.1 Chi-Squared Goodness-of-Fit Tests, 212
 - 7.2.2 Loglinear Cell Residuals, 213
 - 7.2.3 Tests about Conditional Associations, 214
 - 7.2.4 Confidence Intervals for Conditional Odds Ratios, 214
 - 7.2.5 Loglinear Models for Higher Dimensions, 215
 - 7.2.6 Example: Automobile Accidents and Seat Belts, 215
 - 7.2.7 Three-Factor Interaction, 218
 - 7.2.8 Large Samples and Statistical vs Practical Significance, 218
- 7.3 The Loglinear–Logistic Connection, 219
 - 7.3.1 Using Logistic Models to Interpret Loglinear Models, 219
 - 7.3.2 Example: Auto Accident Data Revisited, 220
 - 7.3.3 Correspondence Between Loglinear and Logistic Models, 221
 - 7.3.4 Strategies in Model Selection, 221
- 7.4 Independence Graphs and Collapsibility, 223
 - 7.4.1 Independence Graphs, 223
 - 7.4.2 Collapsibility Conditions for Three-Way Tables, 224
 - 7.4.3 Collapsibility and Logistic Models, 225
 - 7.4.4 Collapsibility and Independence Graphs for Multiway Tables, 225
 - 7.4.5 Example: Model Building for Student Drug Use, 226
 - 7.4.6 Graphical Models, 228

- 7.5 Modeling Ordinal Associations, 228
 - 7.5.1 Linear-by-Linear Association Model, 229
 - 7.5.2 Example: Sex Opinions, 230
 - 7.5.3 Ordinal Tests of Independence, 232
- Problems, 232

8. Models for Matched Pairs 244

- 8.1 Comparing Dependent Proportions, 245
 - 8.1.1 McNemar Test Comparing Marginal Proportions, 245
 - 8.1.2 Estimating Differences of Proportions, 246
- 8.2 Logistic Regression for Matched Pairs, 247
 - 8.2.1 Marginal Models for Marginal Proportions, 247
 - 8.2.2 Subject-Specific and Population-Averaged Tables, 248
 - 8.2.3 Conditional Logistic Regression for Matched-Pairs, 249
 - 8.2.4 Logistic Regression for Matched Case–Control Studies, 250
 - 8.2.5 Connection between McNemar and Cochran–Mantel–Haenszel Tests, 252
- 8.3 Comparing Margins of Square Contingency Tables, 252
 - 8.3.1 Marginal Homogeneity and Nominal Classifications, 253
 - 8.3.2 Example: Coffee Brand Market Share, 253
 - 8.3.3 Marginal Homogeneity and Ordered Categories, 254
 - 8.3.4 Example: Recycle or Drive Less to Help Environment?, 255
- 8.4 Symmetry and Quasi-Symmetry Models for Square Tables, 256
 - 8.4.1 Symmetry as a Logistic Model, 257
 - 8.4.2 Quasi-Symmetry, 257
 - 8.4.3 Example: Coffee Brand Market Share Revisited, 257
 - 8.4.4 Testing Marginal Homogeneity Using Symmetry and Quasi-Symmetry, 258
 - 8.4.5 An Ordinal Quasi-Symmetry Model, 258
 - 8.4.6 Example: Recycle or Drive Less?, 259
 - 8.4.7 Testing Marginal Homogeneity Using Symmetry and Ordinal Quasi-Symmetry, 259
- 8.5 Analyzing Rater Agreement, 260
 - 8.5.1 Cell Residuals for Independence Model, 261
 - 8.5.2 Quasi-independence Model, 261
 - 8.5.3 Odds Ratios Summarizing Agreement, 262
 - 8.5.4 Quasi-Symmetry and Agreement Modeling, 263
 - 8.5.5 Kappa Measure of Agreement, 264
- 8.6 Bradley–Terry Model for Paired Preferences, 264
 - 8.6.1 The Bradley–Terry Model, 265
 - 8.6.2 Example: Ranking Men Tennis Players, 265
- Problems, 266

9. Modeling Correlated, Clustered Responses 276

- 9.1 Marginal Models Versus Conditional Models, 277
 - 9.1.1 Marginal Models for a Clustered Binary Response, 277
 - 9.1.2 Example: Longitudinal Study of Treatments for Depression, 277
 - 9.1.3 Conditional Models for a Repeated Response, 279
 - 9.2 Marginal Modeling: The GEE Approach, 279
 - 9.2.1 Quasi-Likelihood Methods, 280
 - 9.2.2 Generalized Estimating Equation Methodology: Basic Ideas, 280
 - 9.2.3 GEE for Binary Data: Depression Study, 281
 - 9.2.4 Example: Teratology Overdispersion, 283
 - 9.2.5 Limitations of GEE Compared with ML, 284
 - 9.3 Extending GEE: Multinomial Responses, 285
 - 9.3.1 Marginal Modeling of a Clustered Multinomial Response, 285
 - 9.3.2 Example: Insomnia Study, 285
 - 9.3.3 Another Way of Modeling Association with GEE, 287
 - 9.3.4 Dealing with Missing Data, 287
 - 9.4 Transitional Modeling, Given the Past, 288
 - 9.4.1 Transitional Models with Explanatory Variables, 288
 - 9.4.2 Example: Respiratory Illness and Maternal Smoking, 288
 - 9.4.3 Comparisons that Control for Initial Response, 289
 - 9.4.4 Transitional Models Relate to Loglinear Models, 290
- Problems, 290

10. Random Effects: Generalized Linear Mixed Models 297

- 10.1 Random Effects Modeling of Clustered Categorical Data, 297
 - 10.1.1 The Generalized Linear Mixed Model, 298
 - 10.1.2 A Logistic GLMM for Binary Matched Pairs, 299
 - 10.1.3 Example: Sacrifices for the Environment Revisited, 300
 - 10.1.4 Differing Effects in Conditional Models and Marginal Models, 300
- 10.2 Examples of Random Effects Models for Binary Data, 302
 - 10.2.1 Small-Area Estimation of Binomial Probabilities, 302
 - 10.2.2 Example: Estimating Basketball Free Throw Success, 303
 - 10.2.3 Example: Teratology Overdispersion Revisited, 304
 - 10.2.4 Example: Repeated Responses on Similar Survey Items, 305
 - 10.2.5 Item Response Models: The Rasch Model, 307
 - 10.2.6 Example: Depression Study Revisited, 307
 - 10.2.7 Choosing Marginal or Conditional Models, 308
 - 10.2.8 Conditional Models: Random Effects Versus Conditional ML, 309

10.3	Extensions to Multinomial Responses or Multiple Random Effect Terms, 310	
10.3.1	Example: Insomnia Study Revisited, 310	
10.3.2	Bivariate Random Effects and Association Heterogeneity, 311	
10.4	Multilevel (Hierarchical) Models, 313	
10.4.1	Example: Two-Level Model for Student Advancement, 314	
10.4.2	Example: Grade Retention, 315	
10.5	Model Fitting and Inference for GLMMS, 316	
10.5.1	Fitting GLMMs, 316	
10.5.2	Inference for Model Parameters and Prediction, 317	
	Problems, 318	
11.	A Historical Tour of Categorical Data Analysis	325
11.1	The Pearson–Yule Association Controversy, 325	
11.2	R. A. Fisher's Contributions, 326	
11.3	Logistic Regression, 328	
11.4	Multiway Contingency Tables and Loglinear Models, 329	
11.5	Final Comments, 331	
	Appendix A: Software for Categorical Data Analysis	332
	Appendix B: Chi-Squared Distribution Values	343
	Bibliography	344
	Index of Examples	346
	Subject Index	350
	Brief Solutions to Some Odd-Numbered Problems	357