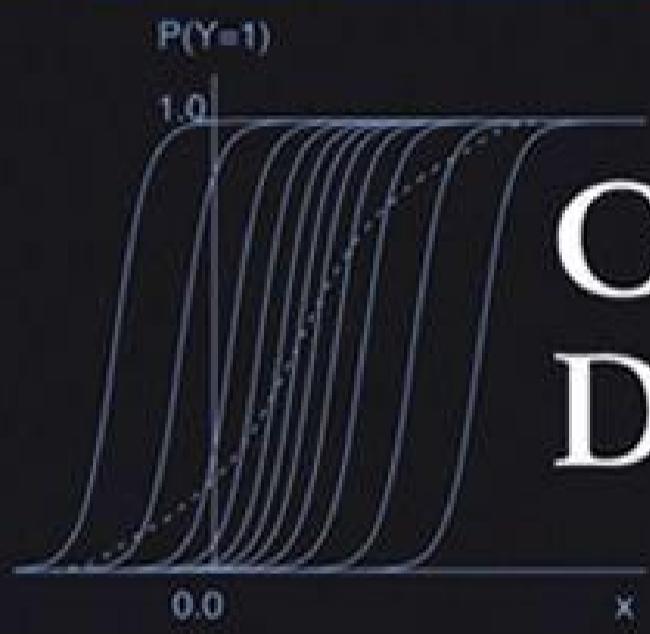


Wiley Series in Probability and Statistics



Categorical Data Analysis

Third Edition

ALAN AGRESTI

 WILEY

WWW.
WILEY.COM

Contents

[Half Title page](#)

[Title page](#)

[Copyright page](#)

[Dedication](#)

[Preface](#)

[Chapter 1: Introduction: Distributions and Inference for Categorical Data](#)

[1.1 Categorical Response Data](#)

[1.2 Distributions for Categorical Data](#)

[1.3 Statistical Inference for Categorical Data](#)

[1.4 Statistical Inference for Binomial Parameters](#)

[1.5 Statistical Inference for Multinomial Parameters](#)

[1.6 Bayesian Inference for Binomial and Multinomial Parameters](#)

[Notes](#)

[Exercises](#)

[Chapter 2: Describing Contingency Tables](#)

[2.1 Probability Structure for Contingency Tables](#)

[2.2 Comparing Two Proportions](#)

[2.3 Conditional Association in Stratified \$2 \times 2\$ Tables](#)

[2.4 Measuring Association in \$I \times J\$ Tables](#)

[Notes](#)

[Exercises](#)

[Chapter 3: Inference for Two-Way Contingency Tables](#)

[3.1 Confidence Intervals for Association Parameters](#)

[3.2 Testing Independence in Two-way Contingency Tables](#)

[3.3 Following-up Chi-Squared Tests](#)

[3.4 Two-Way Tables with Ordered Classifications](#)

[3.5 Small-Sample Inference for Contingency Tables](#)

[3.6 Bayesian Inference for Two-way Contingency Tables](#)

[3.7 Extensions for Multiway Tables and Nontabulated Responses](#)

[Notes](#)

[Exercises](#)

Chapter 4: Introduction to Generalized Linear Models

- [4.1 The Generalized Linear Model](#)
- [4.2 Generalized Linear Models for Binary Data](#)
- [4.3 Generalized Linear Models for Counts and Rates](#)
- [4.4 Moments and Likelihood for Generalized Linear Models](#)
- [4.5 Inference and Model Checking for Generalized Linear Models](#)
- [4.6 Fitting Generalized Linear Models](#)
- [4.7 Quasi-Likelihood and Generalized Linear Models](#)

[Notes](#)

[Exercises](#)

Chapter 5: Logistic Regression

- [5.1 Interpreting Parameters in Logistic Regression](#)
- [5.2 Inference for Logistic Regression](#)
- [5.3 Logistic Models with Categorical Predictors](#)
- [5.4 Multiple Logistic Regression](#)
- [5.5 Fitting Logistic Regression Models](#)

[Notes](#)

[Exercises](#)

Chapter 6: Building, Checking, and Applying Logistic Regression Models

- [6.1 Strategies in Model Selection](#)
- [6.2 Logistic Regression Diagnostics](#)
- [6.3 Summarizing the Predictive Power of a Model](#)
- [6.4 Mantel–Haenszel and Related Methods for Multiple \$2 \times 2\$ Tables](#)
- [6.5 Detecting and Dealing with Infinite Estimates](#)
- [6.6 Sample Size and Power Considerations](#)

[Notes](#)

[Exercises](#)

Chapter 7: Alternative Modeling of Binary Response Data

- [7.1 Probit and Complementary Log–log Models](#)
- [7.2 Bayesian Inference for Binary Regression](#)
- [7.3 Conditional Logistic Regression](#)
- [7.4 Smoothing: Kernels, Penalized Likelihood, Generalized Additive Models](#)
- [7.5 Issues in Analyzing High-Dimensional Categorical Data](#)

[Notes](#)

[Exercises](#)

Chapter 8: Models for Multinomial Responses

- [8.1 Nominal Responses: Baseline-Category Logit Models](#)

[8.2 Ordinal Responses: Cumulative Logit Models](#)
[8.3 Ordinal Responses: Alternative Models](#)
[8.4 Testing Conditional Independence in \$I \times J \times K\$ Tables](#)
[8.5 Discrete-Choice Models](#)
[8.6 Bayesian Modeling of Multinomial Responses](#)
[Notes](#)
[Exercises](#)

[Chapter 9: Loglinear Models for Contingency Tables](#)

[9.1 Loglinear Models for Two-way Tables](#)
[9.2 Loglinear Models for Independence and Interaction in Three-way Tables](#)
[9.3 Inference for Loglinear Models](#)
[9.4 Loglinear Models for Higher Dimensions](#)
[9.5 Loglinear—Logistic Model Connection](#)
[9.6 Loglinear Model Fitting: Likelihood Equations and Asymptotic Distributions](#)
[9.7 Loglinear Model Fitting: Iterative Methods and Their Application](#)
[Notes](#)
[Exercises](#)

[Chapter 10: Building and Extending Loglinear Models](#)

[10.1 Conditional Independence Graphs and Collapsibility](#)
[10.2 Model Selection and Comparison](#)
[10.3 Residuals for Detecting Cell-Specific Lack of Fit](#)
[10.4 Modeling Ordinal Associations](#)
[10.5 Generalized Loglinear and Association Models, Correlation Models, and Correspondence Analysis](#)
[10.6 Empty Cells and Sparseness in Modeling Contingency Tables](#)
[10.7 Bayesian Loglinear Modeling](#)
[Notes](#)
[Exercises](#)

[Chapter 11: Models for Matched Pairs](#)

[11.1 Comparing Dependent Proportions](#)
[11.2 Conditional Logistic Regression for Binary Matched Pairs](#)
[11.3 Marginal Models for Square Contingency Tables](#)
[11.4 Symmetry, Quasi-Symmetry, and Quasi-Independence](#)
[11.5 Measuring Agreement Between Observers](#)
[11.6 Bradley–Terry Model for Paired Preferences](#)
[11.7 Marginal Models and Quasi-Symmetry Models for Matched Sets](#)
[Notes](#)
[Exercises](#)

[Chapter 12: Clustered Categorical Data: Marginal and](#)

Transitional Models

- [12.1 Marginal Modeling: Maximum Likelihood Approach](#)
- [12.2 Marginal Modeling: Generalized Estimating Equations \(GEEs\) Approach](#)
- [12.3 Quasi-Likelihood and Its GEE Multivariate Extension: Details](#)
- [12.4 Transitional Models: Markov Chain and Time Series Models](#)
- [Notes](#)
- [Exercises](#)

Chapter 13: Clustered Categorical Data: Random Effects Models

- [13.1 Random Effects Modeling of Clustered Categorical Data](#)
- [13.2 Binary Responses: Logistic-Normal Model](#)
- [13.3 Examples of Random Effects Models for Binary Data](#)
- [13.4 Random Effects Models for Multinomial Data](#)
- [13.5 Multilevel Modeling](#)
- [13.6 GLMM Fitting, Inference, and Prediction](#)
- [13.7 Bayesian Multivariate Categorical Modeling](#)
- [Notes](#)
- [Exercises](#)

Chapter 14: Other Mixture Models for Discrete Data

- [14.1 Latent Class Models](#)
- [14.2 Nonparametric Random Effects Models](#)
- [14.3 Beta-Binomial Models](#)
- [14.4 Negative Binomial Regression](#)
- [14.5 Poisson Regression with Random Effects](#)
- [Notes](#)
- [Exercises](#)

Chapter 15: Non-Model-Based Classification and Clustering

- [15.1 Classification: Linear Discriminant Analysis](#)
- [15.2 Classification: Tree-Structured Prediction](#)
- [15.3 Cluster Analysis for Categorical Data](#)
- [Notes](#)
- [Exercises](#)

Chapter 16: Large- and Small-Sample Theory for Multinomial Models

- [16.1 Delta Method](#)
- [16.2 Asymptotic Distributions of Estimators of Model Parameters and Cell Probabilities](#)
- [16.3 Asymptotic Distributions of Residuals and Goodness-of-fit Statistics](#)
- [16.4 Asymptotic Distributions for Logit/Loglinear Models](#)

[16.5 Small-Sample Significance Tests for Contingency Tables](#)

[16.6 Small-Sample Confidence Intervals for Categorical Data](#)

[16.7 Alternative Estimation Theory for Parametric Models](#)

[Notes](#)

[Exercises](#)

Chapter 17: Historical Tour of Categorical Data Analysis

[17.1 Pearson–Yule Association Controversy](#)

[17.2 R. A. Fisher’s Contributions](#)

[17.3 Logistic Regression](#)

[17.4 Multiway Contingency Tables and Loglinear Models](#)

[17.5 Bayesian Methods for Categorical Data](#)

[17.6 A Look Forward, and Backward](#)

Appendix A: Statistical Software for Categorical Data Analysis

[A.1 SAS](#)

[A.2 R And S-Plus](#)

[A.3 Stata](#)

[A.4 SPSS](#)

[A.5 Statxact and Logxact](#)

[A.6 Other Software](#)

Appendix B: Chi-Squared Distribution Values

References

Author Index

Example Index

Subject Index

Categorical Data Analysis

WILEY SERIES IN PROBABILITY AND STATISTICS

ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Harvey Goldstein, Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg*

Editors Emeriti: *Vic Barnett, J. Stuart Hunter, Joseph B. Kadane, Jozef L. Teugels*

The **Wiley Series in Probability and Statistics** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

† ABRAHAM and LEDOLTER • Statistical Methods for Forecasting

AGRESTI • Analysis of Ordinal Categorical Data, *Second Edition*

AGRESTI • An Introduction to Categorical Data Analysis, *Second Edition*

AGRESTI • Categorical Data Analysis, *Third Edition*

ALTMAN, GILL, and McDONALD • Numerical Issues in Statistical Computing for the Social Scientist

AMARATUNGA and CABRERA • Exploration and Analysis of DNA Microarray and Protein Array Data

ANDĚL • Mathematics of Chance

ANDERSON • An Introduction to Multivariate Statistical Analysis, *Third Edition*

* ANDERSON • The Statistical Analysis of Time Series

ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG • Statistical Methods for Comparative Studies

ANDERSON and LOYNES • The Teaching of Practical Statistics

ARMITAGE and DAVID (editors) • Advances in Biometry

ARNOLD, BALAKRISHNAN, and NAGARAJA • Records

* ARTHANARI and DODGE • Mathematical Programming in Statistics

* BAILEY • The Elements of Stochastic Processes with Applications to the Natural Sciences

BAJORSKI • Statistics for Imaging, Optics, and Photonics

BALAKRISHNAN and KOUTRAS • Runs and Scans with Applications

BALAKRISHNAN and NG • Precedence-Type Tests and Applications

BARNETT • Comparative Statistical Inference, *Third Edition*

BARNETT • Environmental Statistics

BARNETT and LEWIS • Outliers in Statistical Data, *Third Edition*

BARTHOLOMEW, KNOTT, and MOUSTAKI • Latent Variable Models and Factor Analysis: A Unified Approach, *Third Edition*

BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ • Probability and Statistical Inference, *Second Edition*

BASILEVSKY • Statistical Factor Analysis and Related Methods: Theory and Applications

BATES and WATTS • Nonlinear Regression Analysis and Its Applications

BECHHOFER, SANTNER, and GOLDSMAN • Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons

- BEIRLANT, GOEGEBEUR, SEGERS, TEUGELS, and DE WAAL • Statistics of Extremes: Theory and Applications
- BELSLEY • Conditioning Diagnostics: Collinearity and Weak Data in Regression
- [†] BELSLEY, KUH, and WELSCH • Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
- BENDAT and PIERSOL • Random Data: Analysis and Measurement Procedures, *Fourth Edition*
- BERNARDO and SMITH • Bayesian Theory
- BERZUNI, DAWID, and BERNARDINELL • Causality: Statistical Perspectives and Applications
- BHAT and MILLER • Elements of Applied Stochastic Processes, *Third Edition*
- BHATTACHARYA and WAYMIRE • Stochastic Processes with Applications
- BIEMER, GROVES, LYBERG, MATHIOWETZ, and SUDMAN • Measurement Errors in Surveys
- BILLINGSLEY • Convergence of Probability Measures, *Second Edition*
- BILLINGSLEY • Probability and Measure, *Anniversary Edition*
- BIRKES and DODGE • Alternative Methods of Regression
- BISGAARD and KULAHCI • Time Series Analysis and Forecasting by Example
- BISWAS, DATTA, FINE, and SEGAL • Statistical Advances in the Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics
- BLISCHKE and MURTHY (editors) • Case Studies in Reliability and Maintenance
- BLISCHKE and MURTHY • Reliability: Modeling, Prediction, and Optimization
- BLOOMFIELD • Fourier Analysis of Time Series: An Introduction, *Second Edition*
- BOLLEN • Structural Equations with Latent Variables
- BOLLEN and CURRAN • Latent Curve Models: A Structural Equation Perspective
- BOROVKOV • Ergodicity and Stability of Stochastic Processes
- BOSQ and BLANKE • Inference and Prediction in Large Dimensions
- BOULEAU • Numerical Methods for Stochastic Processes
- * BOX and TIAO • Bayesian Inference in Statistical Analysis
- BOX • Improving Almost Anything, *Revised Edition*
- * BOX and DRAPER • Evolutionary Operation: A Statistical Method for Process Improvement
- BOX and DRAPER • Response Surfaces, Mixtures, and Ridge Analyses, *Second Edition*
- BOX, HUNTER, and HUNTER • Statistics for Experimenters: Design, Innovation, and Discovery, *Second Edition*
- BOX, JENKINS, and REINSEL • Time Series Analysis: Forecasting and Control, *Fourth Edition*
- BOX, LUCEÑO, and PANIAGUA-QUIÑONES • Statistical Control by Monitoring and Adjustment, *Second Edition*
- * BROWN and HOLLANDER • Statistics: A Biomedical Introduction
- CAIROLI and DALANG • Sequential Stochastic Optimization
- CASTILLO, HADI, BALAKRISHNAN, and SARABIA • Extreme Value and Related Models with Applications in Engineering and Science
- CHAN • Time Series: Applications to Finance with R and S-Plus®, *Second Edition*
- CHARALAMBIDES • Combinatorial Methods in Discrete Distributions
- CHATTERJEE and HADI • Regression Analysis by Example, *Fifth Edition*
- CHATTERJEE and HADI • Sensitivity Analysis in Linear Regression
- CHERNICK • Bootstrap Methods: A Guide for Practitioners and Researchers, *Second Edition*
- CHERNICK and FRIIS • Introductory Biostatistics for the Health Sciences
- CHILÈS and DELFINER • Geostatistics: Modeling Spatial Uncertainty, *Second Edition*
- CHOW and LIU • Design and Analysis of Clinical Trials: Concepts and Methodologies, *Second Edition*
- CLARKE • Linear Models: The Theory and Application of Analysis of Variance
- CLARKE and DISNEY • Probability and Random Processes: A First Course with Applications,

Second Edition

* COCHRAN and COX • Experimental Designs, *Second Edition*

COLLINS and LANZA • Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences

CONGDON • Applied Bayesian Modelling

CONGDON • Bayesian Models for Categorical Data

CONGDON • Bayesian Statistical Modelling, *Second Edition*

CONOVER • Practical Nonparametric Statistics, *Third Edition*

COOK • Regression Graphics

COOK and WEISBERG • An Introduction to Regression Graphics

COOK and WEISBERG • Applied Regression Including Computing and Graphics

CORNELL • A Primer on Experiments with Mixtures

CORNELL • Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Third Edition*

COX • A Handbook of Introductory Statistical Methods

CRESSIE • Statistics for Spatial Data, *Revised Edition*

CRESSIE and WIKLE • Statistics for Spatio-Temporal Data

CSÖRGÖ and HORVÁTH • Limit Theorems in Change Point Analysis

DAGPUNAR • Simulation and Monte Carlo: With Applications in Finance and MCMC

DANIEL • Applications of Statistics to Industrial Experimentation

DANIEL • Biostatistics: A Foundation for Analysis in the Health Sciences, *Eighth Edition*

* DANIEL • Fitting Equations to Data: Computer Analysis of Multifactor Data, *Second Edition*

DASU and JOHNSON • Exploratory Data Mining and Data Cleaning

DAVID and NAGARAJA • Order Statistics, *Third Edition*

* DEGROOT, FIENBERG, and KADANE • Statistics and the Law

DEL CASTILLO • Statistical Process Adjustment for Quality Control

DEMARIS • Regression with Social Data: Modeling Continuous and Limited Response Variables

DEMIDENKO • Mixed Models: Theory and Applications

DENISON, HOLMES, MALLICK and SMITH • Bayesian Methods for Nonlinear Classification and Regression

DETTE and STUDDEN • The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis

DEY and MUKERJEE • Fractional Factorial Plans

DILLON and GOLDSTEIN • Multivariate Analysis: Methods and Applications

* DODGE and ROMIG • Sampling Inspection Tables, *Second Edition*

* DOOB • Stochastic Processes

DOWDY, WEARDEN, and CHILKO • Statistics for Research, *Third Edition*

DRAPER and SMITH • Applied Regression Analysis, *Third Edition*

DRYDEN and MARDIA • Statistical Shape Analysis

DUDEWICZ and MISHRA • Modern Mathematical Statistics

DUNN and CLARK • Basic Statistics: A Primer for the Biomedical Sciences, *Fourth Edition*

DUPUIS and ELLIS • A Weak Convergence Approach to the Theory of Large Deviations

EDLER and KITSOS • Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment

* ELANDT-JOHNSON and JOHNSON • Survival Models and Data Analysis

ENDERS • Applied Econometric Time Series, *Third Edition*

† ETHIER and KURTZ • Markov Processes: Characterization and Convergence

EVANS, HASTINGS, and PEACOCK • Statistical Distributions, *Third Edition*

EVERITT, LANDAU, LEESE, and STAHL • Cluster Analysis, *Fifth Edition*

- FEDERER and KING • Variations on Split Plot and Split Block Experiment Designs
- FELLER • An Introduction to Probability Theory and Its Applications, Volume I, *Third Edition*, Revised; Volume II, *Second Edition*
- FITZMAURICE, LAIRD, and WARE • Applied Longitudinal Analysis, *Second Edition*
- * FLEISS • The Design and Analysis of Clinical Experiments
- FLEISS • Statistical Methods for Rates and Proportions, *Third Edition*
- † FLEMING and HARRINGTON • Counting Processes and Survival Analysis
- FUJIKOSHI, ULYANOV, and SHIMIZU • Multivariate Statistics: High-Dimensional and Large-Sample Approximations
- FULLER • Introduction to Statistical Time Series, *Second Edition*
- ‡ FULLER • Measurement Error Models
- GALLANT • Nonlinear Statistical Models
- GEISSER • Modes of Parametric Statistical Inference
- GELMAN and MENG • Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives
- GEWEKE • Contemporary Bayesian Econometrics and Statistics
- GHOSH, MUKHOPADHYAY, and SEN • Sequential Estimation
- GIESBRECHT and GUMPERTZ • Planning, Construction, and Statistical Analysis of Comparative Experiments
- GIFI • Nonlinear Multivariate Analysis
- GIVENS and HOETING • Computational Statistics
- GLASSERMAN and YAO • Monotone Structure in Discrete-Event Systems
- GNANADESIKAN • Methods for Statistical Data Analysis of Multivariate Observations, *Second Edition*
- GOLDSTEIN • Multilevel Statistical Models, *Fourth Edition*
- GOLDSTEIN and LEWIS • Assessment: Problems, Development, and Statistical Issues
- GOLDSTEIN and WOOFF • Bayes Linear Statistics
- GREENWOOD and NIKULIN • A Guide to Chi-Squared Testing
- GROSS, SHORTLE, THOMPSON, and HARRIS • Fundamentals of Queueing Theory, *Fourth Edition*
- GROSS, SHORTLE, THOMPSON, and HARRIS • Solutions Manual to Accompany Fundamentals of Queueing Theory, *Fourth Edition*
- * HAHN and SHAPIRO • Statistical Models in Engineering
- HAHN and MEEKER • Statistical Intervals: A Guide for Practitioners
- HALD • A History of Probability and Statistics and their Applications Before 1750
- † HAMPEL • Robust Statistics: The Approach Based on Influence Functions
- HARTUNG, KNAPP, and SINHA • Statistical Meta-Analysis with Applications
- HEIBERGER • Computation for the Analysis of Designed Experiments
- HEDAYAT and SINHA • Design and Inference in Finite Population Sampling
- HEDEKER and GIBBONS • Longitudinal Data Analysis
- HELLER • MACSYMA for Statisticians
- HERITIER, CANTONI, COPT, and VICTORIA-FESER • Robust Methods in Biostatistics
- HINKELMANN and KEMPTHORNE • Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design, *Second Edition*
- HINKELMANN and KEMPTHORNE • Design and Analysis of Experiments, Volume 2: Advanced Experimental Design
- HINKELMANN (editor) • Design and Analysis of Experiments, Volume 3: Special Designs and Applications
- HOAGLIN, MOSTELLER, and TUKEY • Fundamentals of Exploratory Analysis of Variance

- * HOAGLIN, MOSTELLER, and TUKEY • Exploring Data Tables, Trends and Shapes
- * HOAGLIN, MOSTELLER, and TUKEY • Understanding Robust and Exploratory Data Analysis
- HOCHBERG and TAMHANE • Multiple Comparison Procedures
- HOCKING • Methods and Applications of Linear Models: Regression and the Analysis of Variance, *Second Edition*
- HOEL • Introduction to Mathematical Statistics, *Fifth Edition*
- HOGG and KLUGMAN • Loss Distributions
- HOLLANDER and WOLFE • Nonparametric Statistical Methods, *Second Edition*
- HOSMER and LEMESHOW • Applied Logistic Regression, *Second Edition*
- HOSMER, LEMESHOW, and MAY • Applied Survival Analysis: Regression Modeling of Time-to-Event Data, *Second Edition*
- HUBER • Data Analysis: What Can Be Learned From the Past 50 Years
- HUBER • Robust Statistics
- † HUBER and RONCHETTI • Robust Statistics, *Second Edition*
- HUBERTY • Applied Discriminant Analysis, *Second Edition*
- HUBERTY and OLEJNIK • Applied MANOVA and Discriminant Analysis, *Second Edition*
- HUITEMA • The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, *Second Edition*
- HUNT and KENNEDY • Financial Derivatives in Theory and Practice, *Revised Edition*
- HURD and MIAMEE • Periodically Correlated Random Sequences: Spectral Theory and Practice
- HUSKOVA, BERAN, and DUPAC • Collected Works of Jaroslav Hajek—with Commentary
- HUZURBAZAR • Flowgraph Models for Multistate Time-to-Event Data
- JACKMAN • Bayesian Analysis for the Social Sciences
- † JACKSON • A User's Guide to Principle Components
- JOHN • Statistical Methods in Engineering and Quality Assurance
- JOHNSON • Multivariate Statistical Simulation
- JOHNSON and BALAKRISHNAN • Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz
- JOHNSON, KEMP, and KOTZ • Univariate Discrete Distributions, *Third Edition*
- JOHNSON and KOTZ (editors) • Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present
- JOHNSON, KOTZ, and BALAKRISHNAN • Continuous Univariate Distributions, Volume 1, *Second Edition*
- JOHNSON, KOTZ, and BALAKRISHNAN • Continuous Univariate Distributions, Volume 2, *Second Edition*
- JOHNSON, KOTZ, and BALAKRISHNAN • Discrete Multivariate Distributions
- JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE • The Theory and Practice of Econometrics, *Second Edition*
- JUREK and MASON • Operator-Limit Distributions in Probability Theory
- KADANE • Bayesian Methods and Ethics in a Clinical Trial Design
- KADANE AND SCHUM • A Probabilistic Analysis of the Sacco and Vanzetti Evidence
- KALBFLEISCH and PRENTICE • The Statistical Analysis of Failure Time Data, *Second Edition*
- KARIYA and KURATA • Generalized Least Squares
- KASS and VOS • Geometrical Foundations of Asymptotic Inference
- † KAUFMAN and ROUSSEEUW • Finding Groups in Data: An Introduction to Cluster Analysis
- KEDEM and FOKIANOS • Regression Models for Time Series Analysis
- KENDALL, BARDEN, CARNE, and LE • Shape and Shape Theory
- KHURI • Advanced Calculus with Applications in Statistics, *Second Edition*
- KHURI, MATHEW, and SINHA • Statistical Tests for Mixed Linear Models

* KISH • Statistical Design for Research

KLEIBER and KOTZ • Statistical Size Distributions in Economics and Actuarial Sciences

KLEMELÄ • Smoothing of Multivariate Data: Density Estimation and Visualization

KLUGMAN, PANJER, and WILLMOT • Loss Models: From Data to Decisions, *Fourth Edition*

KLUGMAN, PANJER, and WILLMOT • Student Solutions Manual to Accompany Loss Models: From Data to Decisions, *Fourth Edition*

KOSKI and NOBLE • Bayesian Networks: An Introduction

KOTZ, BALAKRISHNAN, and JOHNSON • Continuous Multivariate Distributions, Volume 1, *Second Edition*

KOTZ and JOHNSON (editors) • Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index

KOTZ and JOHNSON (editors) • Encyclopedia of Statistical Sciences: Supplement Volume

KOTZ, READ, and BANKS (editors) • Encyclopedia of Statistical Sciences: Update Volume 1

KOTZ, READ, and BANKS (editors) • Encyclopedia of Statistical Sciences: Update Volume 2

KOWALSKI and TU • Modern Applied U-Statistics

KRISHNAMOORTHY and MATHEW • Statistical Tolerance Regions: Theory, Applications, and Computation

KROESE, TAIMRE, and BOTEV • Handbook of Monte Carlo Methods

KROONENBERG • Applied Multiway Data Analysis

KULINSKAYA, MORGENTHALER, and STAUDTE • Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence

KULKARNI and HARMAN • An Elementary Introduction to Statistical Learning Theory

KUROWICKA and COOKE • Uncertainty Analysis with High Dimensional Dependence Modelling

KVAM and VIDAKOVIC • Nonparametric Statistics with Applications to Science and Engineering

LACHIN • Biostatistical Methods: The Assessment of Relative Risks, *Second Edition*

LAD • Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction

LAMPERTI • Probability: A Survey of the Mathematical Theory, *Second Edition*

LAWLESS • Statistical Models and Methods for Lifetime Data, *Second Edition*

LAWSON • Statistical Methods in Spatial Epidemiology, *Second Edition*

LE • Applied Categorical Data Analysis, *Second Edition*

LE • Applied Survival Analysis

LEE • Structural Equation Modeling: A Bayesian Approach

LEE and WANG • Statistical Methods for Survival Data Analysis, *Third Edition*

LEPAGE and BILLARD • Exploring the Limits of Bootstrap

LESSLER and KALSBECK • Nonsampling Errors in Surveys

LEYLAND and GOLDSTEIN (editors) • Multilevel Modelling of Health Statistics

LIAO • Statistical Group Comparison

LIN • Introductory Stochastic Analysis for Finance and Insurance

LITTLE and RUBIN • Statistical Analysis with Missing Data, *Second Edition*

LLOYD • The Statistical Analysis of Categorical Data

LOWEN and TEICH • Fractal-Based Point Processes

MAGNUS and NEUDECKER • Matrix Differential Calculus with Applications in Statistics and Econometrics, *Revised Edition*

MALLER and ZHOU • Survival Analysis with Long Term Survivors

MARCHETTE • Random Graphs for Statistical Pattern Recognition

MARDIA and JUPP • Directional Statistics

MARKOVICH • Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice

MARONNA, MARTIN and YOHAI • Robust Statistics: Theory and Methods

MASON, GUNST, and HESS • Statistical Design and Analysis of Experiments with Applications to

Engineering and Science, *Second Edition*

McCOOL • Using the Weibull Distribution: Reliability, Modeling, and Inference

McCULLOCH, SEARLE, and NEUHAUS • Generalized, Linear, and Mixed Models, *Second Edition*

McFADDEN • Management of Data in Clinical Trials, *Second Edition*

* McLACHLAN • Discriminant Analysis and Statistical Pattern Recognition

McLACHLAN, DO, and AMBROISE • Analyzing Microarray Gene Expression Data

McLACHLAN and KRISHNAN • The EM Algorithm and Extensions, *Second Edition*

McLACHLAN and PEEL • Finite Mixture Models

MCNEIL • Epidemiological Research Methods

MEEKER and ESCOBAR • Statistical Methods for Reliability Data

MEERSCHAERT and SCHEFFLER • Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice

MENGERSEN, ROBERT, and TITTERINGTON • Mixtures: Estimation and Applications

MICKEY, DUNN, and CLARK • Applied Statistics: Analysis of Variance and Regression, *Third Edition*

* MILLER • Survival Analysis, *Second Edition*

MONTGOMERY, JENNINGS, and KULAHCI • Introduction to Time Series Analysis and Forecasting

MONTGOMERY, PECK, and VINING • Introduction to Linear Regression Analysis, *Fifth Edition*

MORGENTHALER and TUKEY • Configural Polysampling: A Route to Practical Robustness

MUIRHEAD • Aspects of Multivariate Statistical Theory

MULLER and STOYAN • Comparison Methods for Stochastic Models and Risks

MURTHY, XIE, and JIANG • Weibull Models

MYERS, MONTGOMERY, and ANDERSON-COOK • Response Surface Methodology: Process and Product Optimization Using Designed Experiments, *Third Edition*

MYERS, MONTGOMERY, VINING, and ROBINSON • Generalized Linear Models. With Applications in Engineering and the Sciences, *Second Edition*

NATVIG • Multistate Systems Reliability Theory With Applications

† NELSON • Accelerated Testing, Statistical Models, Test Plans, and Data Analyses

† NELSON • Applied Life Data Analysis

NEWMAN • Biostatistical Methods in Epidemiology

NG, TAIN, and TANG • Dirichlet Theory: Theory, Methods and Applications

OKABE, BOOTS, SUGIHARA, and CHIU • Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, *Second Edition*

OLIVER and SMITH • Influence Diagrams, Belief Nets and Decision Analysis

PALTA • Quantitative Methods in Population Health: Extensions of Ordinary Regressions

PANJER • Operational Risk: Modeling and Analytics

PANKRATZ • Forecasting with Dynamic Regression Models

PANKRATZ • Forecasting with Univariate Box-Jenkins Models: Concepts and Cases

PARDOUX • Markov Processes and Applications: Algorithms, Networks, Genome and Finance

PARMIGIANI and INOUE • Decision Theory: Principles and Approaches

* PARZEN • Modern Probability Theory and Its Applications

PEÑA, TIAO, and TSAY • A Course in Time Series Analysis

PESARIN and SALMASO • Permutation Tests for Complex Data: Applications and Software

PIANTADOSI • Clinical Trials: A Methodologic Perspective, *Second Edition*

POURAHMADI • Foundations of Time Series Analysis and Prediction Theory

POWELL • Approximate Dynamic Programming: Solving the Curses of Dimensionality, *Second Edition*

POWELL and RYZHOV • Optimal Learning

- PRESS • Subjective and Objective Bayesian Statistics, *Second Edition*
PRESS and TANUR • The Subjectivity of Scientists and the Bayesian Approach
PURI, VILAPLANA, and WERTZ • New Perspectives in Theoretical and Applied Statistics
† PUTERMAN • Markov Decision Processes: Discrete Stochastic Dynamic Programming
QIU • Image Processing and Jump Regression Analysis
* RAO • Linear Statistical Inference and Its Applications, *Second Edition*
RAO • Statistical Inference for Fractional Diffusion Processes
RAUSAND and HØYLAND • System Reliability Theory: Models, Statistical Methods, and Applications, *Second Edition*
RAYNER, THAS, and BEST • Smooth Tests of Goodness of Fit: Using R, *Second Edition*
RENCHER and SCHAALJE • Linear Models in Statistics, *Second Edition*
RENCHER and CHRISTENSEN • Methods of Multivariate Analysis, *Third Edition*
RENCHER • Multivariate Statistical Inference with Applications
RIGDON and BASU • Statistical Methods for the Reliability of Repairable Systems
* RIPLEY • Spatial Statistics
* RIPLEY • Stochastic Simulation
ROHATGI and SALEH • An Introduction to Probability and Statistics, *Second Edition*
ROLSKI, SCHMIDLI, SCHMIDT, and TEUGELS • Stochastic Processes for Insurance and Finance
ROSENBERGER and LACHIN • Randomization in Clinical Trials: Theory and Practice
ROSSI, ALLENBY, and MCCULLOCH • Bayesian Statistics and Marketing
† ROUSSEEUW and LEROY • Robust Regression and Outlier Detection
ROYSTON and SAUERBREI • Multivariate Model Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modeling Continuous Variables
* RUBIN • Multiple Imputation for Nonresponse in Surveys
RUBINSTEIN and KROESE • Simulation and the Monte Carlo Method, *Second Edition*
RUBINSTEIN and MELAMED • Modern Simulation and Modeling
RYAN • Modern Engineering Statistics
RYAN • Modern Experimental Design
RYAN • Modern Regression Methods, *Second Edition*
RYAN • Statistical Methods for Quality Improvement, *Third Edition*
SALEH • Theory of Preliminary Test and Stein-Type Estimation with Applications
SALTELLI, CHAN, and SCOTT (editors) • Sensitivity Analysis
SCHERER • Batch Effects and Noise in Microarray Experiments: Sources and Solutions
* SCHEFFE • The Analysis of Variance
SCHIMEK • Smoothing and Regression: Approaches, Computation, and Application
SCHOTT • Matrix Analysis for Statistics, *Second Edition*
SCHOUTENS • Levy Processes in Finance: Pricing Financial Derivatives
SCOTT • Multivariate Density Estimation: Theory, Practice, and Visualization
* SEARLE • Linear Models
† SEARLE • Linear Models for Unbalanced Data
† SEARLE • Matrix Algebra Useful for Statistics
† SEARLE, CASELLA, and McCULLOCH • Variance Components
SEARLE and WILLETT • Matrix Algebra for Applied Economics
SEBER • A Matrix Handbook For Statisticians
† SEBER • Multivariate Observations
SEBER and LEE • Linear Regression Analysis, *Second Edition*
† SEBER and WILD • Nonlinear Regression
SENNOTT • Stochastic Dynamic Programming and the Control of Queueing Systems

- * SERFLING • Approximation Theorems of Mathematical Statistics
SHAFER and VOVK • Probability and Finance: It's Only a Game!
SHERMAN • Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties
SILVAPULLE and SEN • Constrained Statistical Inference: Inequality, Order, and Shape Restrictions
SINGPURWALLA • Reliability and Risk: A Bayesian Perspective
SMALL and MCLEISH • Hilbert Space Methods in Probability and Statistical Inference
SRIVASTAVA • Methods of Multivariate Statistics
STAPLETON • Linear Statistical Models, *Second Edition*
STAPLETON • Models for Probability and Statistical Inference: Theory and Applications
STAUDTE and SHEATHER • Robust Estimation and Testing
STOYAN • Counterexamples in Probability, *Second Edition*
STOYAN, KENDALL, and MECKE • Stochastic Geometry and Its Applications, *Second Edition*
STOYAN and STOYAN • Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics
STREET and BURGESS • The Construction of Optimal Stated Choice Experiments: Theory and Methods
STYAN • The Collected Papers of T. W. Anderson: 1943–1985
SUTTON, ABRAMS, JONES, SHELDON, and SONG • Methods for Meta-Analysis in Medical Research
TAKEZAWA • Introduction to Nonparametric Regression
TAMHANE • Statistical Analysis of Designed Experiments: Theory and Applications
TANAKA • Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
THOMPSON • Empirical Model Building: Data, Models, and Reality, *Second Edition*
THOMPSON • Sampling, *Third Edition*
THOMPSON • Simulation: A Modeler's Approach
THOMPSON and SEBER • Adaptive Sampling
THOMPSON, WILLIAMS, and FINDLAY • Models for Investors in Real World Markets
TIERNEY • LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
TSAY • Analysis of Financial Time Series, *Third Edition*
TSAY • An Introduction to Analysis of Financial Data with R
UPTON and FINGLETON • Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
† VAN BELLE • Statistical Rules of Thumb, *Second Edition*
VAN BELLE, FISHER, HEAGERTY, and LUMLEY • Biostatistics: A Methodology for the Health Sciences, *Second Edition*
VESTRUP • The Theory of Measures and Integration
VIDAKOVIC • Statistical Modeling by Wavelets
VIERTL • Statistical Methods for Fuzzy Data
VINOD and REAGLE • Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments
WALLER and GOTWAY • Applied Spatial Statistics for Public Health Data
WEISBERG • Applied Linear Regression, *Third Edition*
WEISBERG • Bias and Causation: Models and Judgment for Valid Comparisons
WELSH • Aspects of Statistical Inference
WESTFALL and YOUNG • Resampling-Based Multiple Testing: Examples and Methods for p -Value Adjustment

- * WHITTAKER • Graphical Models in Applied Multivariate Statistics
- WINKER • Optimization Heuristics in Economics: Applications of Threshold Accepting
- WOODWORTH • Biostatistics: A Bayesian Introduction
- WOOLSON and CLARKE • Statistical Methods for the Analysis of Biomedical Data, *Second Edition*
- WU and HAMADA • Experiments: Planning, Analysis, and Parameter Design Optimization, *Second Edition*
- WU and ZHANG • Nonparametric Regression Methods for Longitudinal Data Analysis
- YIN • Clinical Trial Design: Bayesian and Frequentist Adaptive Methods
- YOUNG, VALERO-MORA, and FRIENDLY • Visual Statistics: Seeing Data with Dynamic Interactive Graphics
- ZACKS • Stage-Wise Adaptive Designs
- * ZELLNER • An Introduction to Bayesian Inference in Econometrics
- ZELTERMAN • Discrete Distributions—Applications in the Health Sciences
- ZHOU, OBUCHOWSKI, and MCCLISH • Statistical Methods in Diagnostic Medicine, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

Categorical Data Analysis

Third Edition

ALAN AGRESTI

Department of Statistics
University of Florida
Gainesville, Florida



Copyright © 2013 by John Wiley & Sons. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400, fax 978-750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, 201-748-6011, fax 201-748-6008, or online at <http://www.wiley.com/go/permission>.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at 800-762-2974, outside the United States at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data

Agresti, Alan.

Categorical data analysis / Alan Agresti. – 3rd ed.

p. cm. – (Wiley series in probability and statistics; 792)

Includes bibliographical references and index.

ISBN 978-0-470-46363-5 (hardback)

1. Multivariate analysis. I. Title.

QA278.A353 2013

519.5'35—dc23

2012009792

To Jacki

Preface

The explosion in the development of methods for analyzing categorical data that began in the 1960s has continued apace in recent years. This book provides an overview of these methods, as well as older, now standard, methods. It gives special emphasis to generalized linear modeling techniques, which extend linear model methods for continuous variables, and their extensions for multivariate responses.

OUTLINE OF TOPICS

Chapters 1–10 present the core methods for categorical response variables. Chapters 1–3 cover distributions for categorical responses and traditional methods for two-way contingency tables. Chapters 4–8 introduce logistic regression and related models such as the probit model for binary and multicategory response variables. Chapters 9 and 10 cover loglinear models for contingency tables.

In the past quarter century, a major area of new research has been the development of methods for repeated measurement and other forms of clustered categorical data. Chapters 11–14 present these methods, including marginal models and generalized linear mixed models with random effects. Chapter 15 introduces non-model-based methods for classification and clustering. Chapter 16 presents theoretical foundations as well as alternatives to the maximum likelihood paradigm that this text adopts. Chapter 17 is devoted to a historical overview of the development of the methods. It examines contributions of noted statisticians, such as Pearson and Fisher, whose pioneering efforts—and sometimes vocal debates—broke the ground for this evolution.

Appendices illustrate the use of statistical software for analyzing categorical data. The website for the text, www.stat.ufl.edu/~aa/cda/cda.html, contains an appendix with detailed examples of the use of software (especially R, SAS, and Stata) for performing the analyses in this book, solutions to many of the exercises, extra exercises, and corrections.

CHANGES IN THIS EDITION

Given the explosion of research in the past 50 years on categorical data methods, it is an increasing challenge to write a comprehensive book covering all the commonly used methods. The second edition of this book already exceeded 700 pages. In including much new material without letting the book grow much, I have necessarily had to make compromises in depth and use relatively simple examples. I try to present a broad overview, while presenting bibliographic notes with many references in which the reader can find more details. In attempting to make the book relatively comprehensive while presenting substantive new material, every chapter of the first two editions has been extensively rewritten. The major changes are:

- A new Chapter 7 presents alternative methods for binary response data, including some regularization methods that are becoming popular in this age of massive data sets with enormous numbers of variables.
- A new Chapter 15 introduces non-model-based methods of classification, such as linear discriminant analysis and classification trees, and cluster analysis.
- Many chapters now include a section describing the Bayesian approach for the methods of that chapter. We also have added material (e.g., Sections 6.5 and 7.4) about ways that frequentist methods can deal with awkward situations such as infinite maximum likelihood estimates.
- The use of various software for categorical data methods is discussed at a much expanded website for the text, www.stat.ufl.edu/~aa/cda/cda.html. Examples are shown of the use of R, SAS, and Stata for most of the examples in the text, and there is discussion also about SPSS, StatXact, and other software. That website also contains many of the text's data sets, some of which have only excerpts shown in the text itself, as well as solutions for many

exercises and corrections of errors found in early printings of the book. I recommend that you refer to this appendix (or specialized software manuals) while reading the text, perhaps printing the pages about the software you prefer, as an aid to implementing the methods. This material was placed at the website partly because the text is already so long without it and also because it is then easier to keep the presentation up-to-date.

In this text, I interpret *categorical data analysis* to refer to methods for categorical response variables. For most methods, explanatory variables can be categorical or quantitative, as in ordinary regression. Thus, the focus is intended to be more general than contingency table analysis, although for simplicity of data presentation, most examples use contingency tables. These examples are simplistic, but should help you focus on understanding the methods themselves and make it easier for you to replicate results with your favorite software.

Other special features of the text include:

- More than 100 analyses of data sets.
- About 600 exercises, some directed toward theory and methods and some toward applications and data analysis.
- Notes at the end of each chapter that provide references for recent research and many topics not covered in the text, linked to a bibliography of more than 1200 sources.

INTENDED AUDIENCE AND USE AS A TEXTBOOK

I intend this book to be accessible to the diverse mix of students who take graduate-level courses in categorical data analysis. But I have also written it with practicing statisticians and biostatisticians in mind. I hope it enables them to catch up with recent advances and learn about methods that sometimes receive inadequate attention in the traditional statistics curriculum.

The development of new methods has influenced—and been influenced by—the increasing availability of data sets with categorical responses in the social, behavioral, and biomedical sciences, as well as in public health, genetics, ecology, education, marketing and the financial industry, and industrial quality control. And so, although this book is directed mainly to statisticians and biostatisticians, I also aim for it to be helpful to methodologists in these fields.

Readers should possess a background that includes regression and analysis of variance models, as well as maximum likelihood methods of statistical theory. Those not having much theory background should be able to follow most methodological discussions. Those with mainly applied interests can skip most of Chapter 4 on the theory of generalized linear models and proceed to other chapters. However, the book has a distinctly higher technical level and is more thorough and complete than my lower-level text, *An Introduction to Categorical Data Analysis, Second Edition* (Wiley, 2007).

Today, because of the ubiquity of categorical data in applications, most statistics and biostatistics departments offer courses on categorical data analysis or on generalized linear models with strong emphasis on methods for discrete data. This book can be used as a text for such courses. The material in Chapters 1–6 forms the heart of most courses. There is too much material in this book for a single course, but a one-term course can be based on the following outline:

- Basic contingency table analysis, covering Chapters 1–3, perhaps skipping some tangential sections such as 1.5–7, 1.6, 2.4, 3.4–3.7.
- Logistic regression and related methods for binary data, covering Chapters 4–6, perhaps skipping some tangential sections such as 4.4–4.7 and 6.4–6.6.
- Multinomial response models, covering at least Sections 8.1 and 8.2.
- Matched pairs and clustered data, covering at least Sections 11.1–11.2.

Courses with biostatistical orientation may want to include bits from Chapters 12 and 13 on marginal and random effects models. Courses with social science emphasis may want to include some topics on loglinear modeling from Chapters 9 and 10. Some courses may want to select specialized topics from Chapter 7, such as probit modeling, conditional logistic regression, Bayesian binary data modeling, smoothing, and issues in the analysis of high-dimensional data.

ACKNOWLEDGMENTS

I thank those who commented on parts of the manuscript or provided help of some type. Special thanks to Anna Gottard, David Hoaglin, Maria Kateri, Bernhard Klingenberg, Keli Liu, and Euijung Ryu, who gave insightful comments on some chapters and made many helpful suggestions, and Brett Presnell for his advice and resources about R software and his comments about some of the material. Thanks to people who made suggestions about new material for this edition, including Jonathan Bischof, James Booth, Brian Caffo, Tianxi Cai, Brent Coull, Nicholas Cox, Ralitsa Gueorguieva, Debasish Ghosh, John Henretta, David Hitchcock, Galin Jones, Robert Kushler, Xihong Lin, Jun Liu, Gianfranco Lovison, Giovanni Marchetti, David Olive, Art Owen, Alessandra Petrucci, Michael Radelet, Gerard Scallan, Maura Stokes, Anestis Touloumis, and Ming Yang. Thanks to those who commented on aspects of the second edition, including pointing out errors or typos, such as Pat Altham, Roberto Bertolusso, Nicholas Cox, David Firth, Rene Gonin, David Hoaglin, Harry Khamis, Bernhard Klingenberg, Robert Kushler, Gianfranco Lovison, Theo Nijssse, Richard Reyment, Misha Salganik, William Santo, Laura Thompson, Michael Vock, and Zhongming Yang. Thanks also to Laura Thompson for preparing her very helpful manual on using R and S-Plus for examples in the second edition. Thanks to the many who reviewed material or suggested examples for the first two editions, mentioned in the Prefaces of those editions. Thanks also to Wiley Executive Editor Steve Quigley and Associate Editor Jacqueline Palmieri for their steadfast encouragement and facilitation of this project. Finally, thanks to my wife Jacki Levine for continuing support of all kinds.

ALAN AGRESTI

Gainesville, Florida and Brookline, Massachusetts
February 2012

CHAPTER 1

Introduction: Distributions and Inference for Categorical Data

From helping to assess the value of new medical treatments to evaluating the factors that affect our opinions and behaviors, analysts today are finding myriad uses for categorical data methods. In this book we introduce these methods and the theory behind them.

Statistical methods for categorical responses were late in gaining the level of sophistication achieved early in the twentieth century by methods for continuous responses. Despite influential work around 1900 by the British statistician Karl Pearson, relatively little development of models for categorical responses occurred until the 1960s. In this book we describe the early fundamental work that still has importance today but place primary emphasis on more recent modeling approaches.

1.1 CATEGORICAL RESPONSE DATA

A *categorical variable* has a measurement scale consisting of a set of categories. For instance, political philosophy is often measured as liberal, moderate, or conservative. Diagnoses regarding breast cancer based on a mammogram use the categories normal, benign, probably benign, suspicious, and malignant.

The development of methods for categorical variables was stimulated by the need to analyze data generated in research studies in both the social and biomedical sciences. Categorical scales are pervasive in the social sciences for measuring attitudes and opinions. Categorical scales in biomedical sciences measure outcomes such as whether a medical treatment is successful.

Categorical data are by no means restricted to the social and biomedical sciences. They frequently occur in the behavioral sciences (e.g., type of mental illness, with the categories schizophrenia, depression, neurosis), epidemiology and public health (e.g., contraceptive method at last sexual intercourse, with the categories none, condom, pill, IUD, other), genetics (type of allele inherited by an offspring), botany and zoology (e.g., whether or not a particular organism is observed in a sampled quadrat), education (e.g., whether a student response to an exam question is correct or incorrect), and marketing (e.g., consumer preference among the three leading brands of a product). They even occur in highly quantitative fields such as engineering sciences and industrial quality control. Examples are the classification of items according to whether they conform to certain standards, and subjective evaluation of some characteristic: how soft to the touch a certain fabric is, how good a particular food product tastes, or how easy a worker finds it to perform a certain task.

Categorical variables are of many types. In this section we provide ways of classifying them.

1.1.1 Response–Explanatory Variable Distinction

Statistical analyses distinguish between *response* (or *dependent*) *variables* and *explanatory* (or *independent*) *variables*. This book focuses on methods for categorical response variables. As in ordinary regression modeling, explanatory variables can be any type. For instance, a study might analyze how opinion about whether same-sex marriages should be legal (yes or no) changes according to values of explanatory variables, such as religious affiliation, political ideology, number of years of education, annual income, age, gender, and race.

1.1.2 Binary–Nominal–Ordinal Scale Distinction

Many categorical variables have only two categories. Such variables, for which the two categories are often given the generic labels “success” and “failure,” are called *binary variables*. A major topic of this book is the modeling of binary response variables.

When a categorical variable has more than two categories, we distinguish between two types of categorical scales. Variables having categories without a natural ordering are said to be measured on a *nominal scale* and are called *nominal variables*. Examples are mode of transportation to get to work (automobile, bicycle, bus, subway, walk), favorite type of music (classical, country, folk, jazz, rock), and choice of residence (apartment, condominium, house, other). For nominal variables, the order of listing the categories is irrelevant to the statistical analysis.

Many categorical variables *do* have ordered categories. Such variables are said to be measured on a *ordinal scale* and are called *ordinal variables*. Examples are social class (upper, middle, lower), political philosophy (very liberal, slightly liberal, moderate, slightly conservative, very conservative), patient condition (good, fair, serious, critical), and rating of a movie for Netflix (1 to 5 stars, representing hated it, didn’t like it, liked it, really liked it, loved it). For ordinal variables, distances between categories are unknown. Although a person categorized as very liberal is more liberal than a person categorized as slightly liberal, no numerical value describes *how much more* liberal that person is.

An *interval variable* is one that *does* have numerical distances between any two values. For example, systolic blood pressure level, length of prison term, and annual income are interval variables. For most such variables, it is also possible to compare two values by their ratio, in which case the variable is also called a *ratio variable*.

The way that a variable is measured determines its classification. For example, “education” is only nominal when measured as (public school, private school, home schooling); it is ordinal when measured by highest degree attained, using the categories (none, high school, bachelor’s, master’s, doctorate); it is interval when measured by number of years of education completed, using the integers 0, 1, 2, 3,

A variable’s measurement scale determines which statistical methods are appropriate. It is usually best to apply methods appropriate for the actual scale. In the measurement hierarchy, interval variables are highest, ordinal variables are next, and nominal variables are lowest. Statistical methods for variables of one type can also be used with variables at higher levels but not at lower levels. For instance, statistical methods for nominal variables can be used with ordinal variables by ignoring the ordering of categories. Methods for ordinal variables cannot, however, be used with nominal variables, since their categories have no meaningful ordering. The distinction between ordered and unordered categories is not important for binary variables, because ordinal methods and nominal methods then typically reduce to equivalent methods.

In this book, we present methods for the analysis of binary, nominal, and ordinal variables. The methods also apply to interval variables having a small number of distinct values (e.g., number of times married, number of distinct side effects experienced in taking some drug) or for which the values are grouped into ordered categories (e.g., education measured as ≤ 12 years, > 12 but < 16 years, ≥ 16 years).

1.1.3 Discrete–Continuous Variable Distinction

Variables are classified as *discrete* or *continuous*, according to whether the number of values they can take is countable. Actual measurement of all variables occurs in a discrete manner, due to precision limitations in measuring instruments. The discrete–continuous classification, in practice, distinguishes between variables that take few values and variables that take lots of values. For instance, statisticians often treat discrete interval variables having a large number of values (such as test scores) as continuous, using them in methods for continuous responses.

This book deals with certain types of discretely measured responses: (1) binary variables, (2) nominal variables, (3) ordinal variables, (4) discrete interval variables having relatively few values, and (5) continuous variables grouped into a small number of categories.

1.1.4 Quantitative–Qualitative Variable Distinction

Nominal variables are *qualitative*—distinct categories differ in quality, not in quantity. Interval variables are *quantitative*—distinct levels have differing amounts of the characteristic of interest. The position of ordinal variables in the qualitative–quantitative classification is fuzzy. Analysts often treat them as qualitative, using methods for nominal variables. But in many respects, ordinal variables more closely resemble interval variables than they resemble nominal variables. They possess important quantitative features: Each category has a *greater* or *smaller* magnitude of the characteristic than another category; and although not possible to measure, an underlying continuous variable is often present. The political ideology classification (very liberal, slightly liberal, moderate, slightly conservative, very conservative) crudely measures an inherently continuous characteristic.

Analysts often utilize the quantitative nature of ordinal variables by assigning numerical scores to the categories or assuming an underlying continuous distribution. This requires good judgment and guidance from researchers who use the scale, but it provides benefits in the variety of methods available for data analysis.

1.1.5 Organization of Book and Online Computing Appendix

The models for categorical response variables discussed in this book resemble regression models for continuous response variables; however, they assume binomial or multinomial response distributions instead of normality. One type of model receives special attention—*logistic regression*. Ordinary logistic regression models apply with *binary* responses and assume a binomial distribution. Generalizations of logistic regression apply with multicategory responses and assume a multinomial distribution.

The book has four main units. In the first, Chapters 1 through 3, we summarize descriptive and inferential methods for univariate and bivariate categorical data. These chapters cover discrete distributions, methods of inference, and measures of association for contingency tables. They summarize the non-model-based methods developed prior to about 1960.

In the second and primary unit, Chapters 4 through 10, we introduce models for categorical responses. In Chapter 4 we describe a class of *generalized linear models* having models of this text as special cases. Chapters 5 and 6 cover the most important model for binary responses, logistic regression. Chapter 7 presents alternative methods for binary data, including the probit, Bayesian fitting, and smoothing methods. In Chapter 8 we present generalizations of the logistic regression model for nominal and ordinal multicategory response variables. In Chapters 9 and 10 we introduce the modeling of multivariate categorical response data, in terms of association and interaction patterns among the variables. The models, called *loglinear models*, apply to counts in the table that cross-classifies those responses.

In the third unit. Chapters 11 through 14, we discuss models for handling repeated measurement and other forms of clustered data. In Chapter 11 we present models for a categorical response with matched pairs; these apply, for instance, with a categorical response measured for the same subjects at two times. Chapter 12 covers models for more general types of repeated categorical data, such as longitudinal data from several times with explanatory variables. In Chapter 13 we present a broad class of models, *generalized linear mixed models*, that use random effects to account for dependence with such data. In Chapter 14 further extensions of the models from Chapters 11 through 13 are described, unified by treating the response as having a mixture distribution of some type.

The fourth and final unit has a different nature than the others. In Chapter 15 we consider non-model-based classification and clustering methods. In Chapter 16 we summarize large-sample and small-sample theory for categorical data models. This theory is the basis for behavior of model parameter estimators and goodness-of-fit statistics. Chapter 17 presents a historical overview of the development of categorical data methods.

Maximum likelihood methods receive primary attention throughout the book. Many chapters, however, contain a section presenting corresponding Bayesian methods.

In Appendix A we review software that can perform the analyses in this book. The website www.stat.ufl.edu/~aa/cda/cda.html for this book contains an appendix that gives more information about using R, SAS, Stata, and other software, with sample programs for text examples. In addition, that site has complete data sets for many text examples and exercises, solutions to some exercises, extra exercises, corrections, and links to other useful sites. For instance, a manual prepared by Dr. Laura Thompson provides examples of how to use R and S-Plus for all examples in the second edition of this text, many of which (or very similar ones) are also in this edition.

In the rest of this chapter, we provide background material. In Section 1.2 we review the key distributions for categorical data: the binomial and multinomial, as well as another that is important for discrete data, the Poisson. In Section 1.3 we review the primary mechanisms for statistical inference using maximum likelihood. In Sections 1.4 and 1.5 we illustrate these by presenting significance tests and confidence intervals for binomial and multinomial parameters. In Section 1.6 we introduce Bayesian inference for these parameters.

1.2 DISTRIBUTIONS FOR CATEGORICAL DATA

Inferential data analyses require assumptions about the random mechanism that generated the data. For regression models with continuous responses, the normal distribution plays the central role. In this section we review the three key distributions for categorical responses: *binomial*, *multinomial*, and *Poisson*.

1.2.1 Binomial Distribution

Many applications refer to a fixed number n of binary observations. Let y_1, y_2, \dots, y_n denote observations from n independent and identical trials such that $P(Y_i = 1) = \pi$ and $P(Y_i = 0) = 1 - \pi$. We refer to outcome 1 as “success” and outcome 0 as “failure.” *Identical trials* means that the probability of success π is the same for each trial. *Independent trials* means that the $\{Y_i\}$ are independent random variables. These are often called *Bernoulli trials*. The total number of successes, $Y = \sum_{i=1}^n Y_i$, has the *binomial distribution* with index n and parameter π , denoted by $\text{bin}(n, \pi)$.

The probability mass function for the possible outcomes y for Y is

$$(1.1) \quad p(y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}, \quad y = 0, 1, 2, \dots, n,$$

where the binomial coefficient $\binom{n}{y} = n!/[y!(n-y)!]$. Since $E(Y_i) = E(Y_i^2) = 1 \times \pi + 0 \times (1 - \pi) = \pi$,

$$E(Y_i) = \pi \quad \text{and} \quad \text{var}(Y_i) = \pi(1 - \pi).$$

The binomial distribution for $Y = \sum_i Y_i$ has mean and variance

$$\mu = E(Y) = n\pi \quad \text{and} \quad \sigma^2 = \text{var}(Y) = n\pi(1 - \pi).$$

The skewness is described by $E(Y - \mu)^3/\sigma^3 = (1 - 2\pi)/\sqrt{n\pi(1 - \pi)}$. The distribution is symmetric when $\pi = 0.50$ but becomes increasingly skewed as π moves toward either boundary. The binomial distribution converges to normality as n increases, for fixed π , the approximation being reasonable¹ when $n[\min(\pi, 1 - \pi)]$ is as small as about 5.

There is no guarantee that successive binary observations are independent or identical. Thus, occasionally, we will utilize other distributions. One such case is sampling binary outcomes without replacement from a finite population, such as observations on whether a homework assignment was completed for 10 students sampled from a class of size 20. The *hypergeometric distribution*, studied in Section 3.5.1, is then relevant. In Section 1.2.4 we discuss another case that violates the binomial assumptions.

1.2.2 Multinomial Distribution

Some trials have more than two possible outcomes. Suppose that each of n independent, identical trials can have outcome in any of c categories. Let $y_{ij} = 1$ if trial i has outcome in category j and $y_{ij} = 0$ otherwise. Then $y_i = (y_{i1}, y_{i2}, \dots, y_{ic})$ represents a multinomial trial, with $\sum_j y_{ij} = 1$; for instance, $(0, 0, 1, 0)$ denotes outcome in category 3 of four possible categories. Note that y_{ic} is redundant, being linearly dependent on the others. Let $n_j = \sum_i y_{ij}$ denote the number of trials having outcome in category j . The counts (n_1, n_2, \dots, n_c) have the *multinomial distribution*.

Let $\pi_j = P(Y_{ij} = 1)$ denote the probability of outcome in category j for each trial. The multinomial probability mass function is

$$(1.2) \quad p(n_1, n_2, \dots, n_{c-1}) = \left(\frac{n!}{n_1! n_2! \dots n_{c-1}!} \right) \pi_1^{n_1} \pi_2^{n_2} \dots \pi_{c-1}^{n_{c-1}}.$$

Since $\sum_j n_j = n$, this is $(c - 1)$ -dimensional, with $n_c = n - (n_1 + \dots + n_{c-1})$. The binomial distribution is the special case with $c = 2$.

For the multinomial distribution,

$$(1.3) \quad E(n_j) = n\pi_j, \quad \text{var}(n_j) = n\pi_j(1 - \pi_j), \quad \text{cov}(n_j, n_k) = -n\pi_j\pi_k.$$

We derive the covariance in Section 16.1.4. The marginal distribution of each n_j is binomial.

1.2.3 Poisson Distribution

Sometimes, count data do not result from a fixed number of trials. For instance, if $Y = \text{number of automobile accidents today on motorways in Italy}$, there is no fixed upper bound n for Y (as you are aware if you have driven in Italy!). Since Y must take a nonnegative integer value, its distribution should place its mass on that range. The simplest such distribution is the *Poisson*. Its probabilities depend on a single parameter, the mean μ . The Poisson probability mass function (Poisson 1837, p. 206) is

$$(1.4) \quad p(y) = \frac{e^{-\mu} \mu^y}{y!}, \quad y = 0, 1, 2, \dots$$

It satisfies $E(Y) = \text{var}(Y) = \mu$. It is unimodal with mode equal to the integer part of μ . Its skewness is described by $E(Y - \mu)^3/\sigma^3 = 1/\sqrt{\mu}$. The Poisson distribution approaches normality as μ increases, the normal approximation being quite good when μ is at least about 10.

The Poisson distribution is used for counts of events that occur randomly over time or space, when outcomes in disjoint periods or regions are independent. It also applies as an approximation for the binomial when n is large and π is small, with $\mu = n\pi$. For example, suppose $Y = \text{number of deaths today in auto accidents in Italy}$ (rather than the *number of accidents*). Then, Y has an upper bound. If each of the 50 million people driving in Italy is an independent trial with probability 0.0000003 of dying today in an auto accident, the number of deaths Y is a $\text{bin}(50000000, 0.0000003)$ variate. This is approximately Poisson with $\mu = n\pi = 50000000(0.0000003) = 15$.

A key feature of the Poisson distribution is that its variance equals its mean. Sample counts vary more when their mean is higher. When the mean number of daily fatal accidents equals 15, greater variability occurs from day to day than when the mean equals 2.

1.2.4 Overdispersion

In practice, count observations often exhibit variability exceeding that predicted by the binomial or Poisson. This phenomenon is called *overdispersion*. We assumed above that each person has the same probability each day of dying in a fatal auto accident. More realistically, these probabilities vary from day to day according to the amount of road traffic and weather conditions and vary from person to person according to factors such as the amount of time spent in autos, whether the person wears a seat belt, how much of the driving is at high speeds, gender, and age. Such variation causes fatality counts to display more variation than predicted by the Poisson model.

Suppose that Y is a random variable with variance $\text{var}(Y|\mu)$ for given μ , but μ itself varies because of unmeasured factors such as those just described. Let $\theta = E(\mu)$. Then unconditionally,

$$E(Y) = E[E(Y|\mu)], \quad \text{var}(Y) = E[\text{var}(Y|\mu)] + \text{var}[E(Y|\mu)].$$

When Y is conditionally Poisson (given μ), then $E(Y) = E(\mu) = \theta$ and $\text{var}(Y) = E(\mu) + \text{var}(\mu) = \theta + \text{var}(\mu) > \theta$.

Assuming a Poisson distribution for a count variable is often too simplistic, because of factors that cause overdispersion. The *negative binomial* is a related distribution for count data that has a second parameter and permits the variance to exceed the mean. We introduce it in Section 4.3.4.

Analyses assuming binomial (or multinomial) distributions are also sometimes invalid because of overdispersion. This might happen because the true distribution is a mixture of different binomial distributions, with the parameter varying because of unmeasured variables. To illustrate, suppose that an experiment exposes pregnant mice to a toxin and then after a week observes the number of fetuses in each mouse's litter that show signs of malformation. Let n_i denote the number of fetuses in the litter for mouse i . The pregnant mice also vary according to other factors, such as their weight, overall health, and genetic makeup. Extra variation then occurs because of the variability from litter to litter in the probability π of malformation. The distribution of the number of fetuses per Utter showing malformations might cluster near 0 and near n_i , showing more dispersion than expected for binomial sampling with a single value of π . Overdispersion could also occur when π varies among fetuses in a litter according to some distribution (Exercise 1.17). In Chapters 4, 13, and 14 we introduce methods for data that are overdispersed relative to binomial and Poisson assumptions.

1.2.5 Connection Between Poisson and Multinomial Distributions

For adult residents of Britain who visit France this year, let Y_1 = number who fly there, Y_2 = number who travel there by train without a car (Eurostar), Y_3 = number who travel there by ferry without a car, and Y_4 = number who take a car (by Eurotunnel Shuttle or a ferry). A Poisson model for (Y_1, Y_2, Y_3, Y_4) treats these as independent Poisson random variables, with parameters $(\mu_1, \mu_2, \mu_3, \mu_4)$. The joint probability mass function for $\{Y_i\}$ is the product of the four mass functions of form (1.4). The total $n = \sum_i Y_i$ also has a Poisson distribution, with parameter $\sum_i \mu_i$.

With Poisson sampling the total count n is random rather than fixed. If we assume a Poisson model but condition on n , $\{Y_i\}$ no longer have Poisson distributions, since each Y_i cannot exceed n . Given n , $\{Y_i\}$ are also no longer independent, since the value of one affects the possible range for the others.

For c independent Poisson variates, with $E(Y_i) = \mu_i$, the conditional probability of a set of counts $\{n_i\}$ satisfying $\sum_i Y_i = n$ is

$$\begin{aligned}
 & P[(Y_1 = n_1, Y_2 = n_2, \dots, Y_c = n_c) | \sum_j Y_j = n] \\
 &= \frac{P(Y_1 = n_1, Y_2 = n_2, \dots, Y_c = n_c)}{P(\sum_j Y_j = n)} \\
 (1.5) \quad &= \frac{\prod_i [\exp(-\mu_i) \mu_i^{n_i} / n_i!]}{\exp(-\sum_j \mu_j) (\sum_j \mu_j)^n / n!} = \frac{n!}{\prod_i n_i!} \prod_i \pi_i^{n_i},
 \end{aligned}$$

where $\pi_i = \mu_i / (\sum_j \mu_j)$. This is the multinomial $(n, \{\pi_i\})$ distribution, characterized by the sample size n and the probabilities $\{\pi_i\}$.

Many categorical data analyses assume a multinomial distribution. Such analyses usually have the same inferential results as those of analyses assuming a Poisson distribution, because of the similarity in the likelihood functions.

1.2.6 The Chi-Squared Distribution

Another distribution of fundamental importance for categorical data is the *chi-squared*, not as a distribution for the data but rather as a sampling distribution for many statistics. Because of its importance, we summarize here a few of its properties.

The chi-squared distribution with degrees of freedom denoted by df has mean df , variance $2(\text{df})$, and skewness $\sqrt{8/\text{df}}$. It converges (slowly) to normality as df increases, the approximation being reasonably good when df is at least about 50.

Let Z denote a standard normal random variable (mean 0, variance 1). Then Z^2 has a chi-squared distribution with $\text{df} = 1$. A chi-squared random variable with $\text{df} = v$ has representation $Z_1^2 + \dots + Z_v^2$, where Z_1, \dots, Z_v are independent standard normal variables. Thus, a chi-squared statistic having $\text{df} = v$ has partitionings into independent chi-squared components—for example, into v components each having $\text{df} = 1$. Conversely, the *reproductive property* states that if X_1^2 and X_2^2 are independent chi-squared random variables having degrees of freedom v_1 and v_2 , then $X^2 = X_1^2 + X_2^2$ has a chi-squared distribution with $\text{df} = v_1 + v_2$.

1.3 STATISTICAL INFERENCE FOR CATEGORICAL DATA

In practice, the probability distribution assumed for the response variable has unknown parameter values. In this section we review methods of using sample data to make inferences about the parameters. Sections 1.4 and 1.5 illustrate these methods for binomial and multinomial parameters.

1.3.1 Likelihood Functions and Maximum Likelihood Estimation

In this book we use *maximum likelihood* for parameter estimation. Maximum likelihood estimators have desirable properties: They have large-sample normal distributions; they are asymptotically consistent, converging to the parameter as n increases; and they are asymptotically efficient, producing large-sample standard errors no greater than those from other estimation methods. These results hold under weak regularity conditions, mainly that the number of parameters remains constant as n increases and that the true values of those parameters fall in the interior (rather than on the boundary) of the parameter space.

Given the data, for a chosen probability distribution the *likelihood function* is the probability of those data, treated as a function of the unknown parameter. The maximum likelihood (ML) estimate is the parameter value that maximizes this function. This is the parameter value under which the data observed have the highest probability of occurrence. We denote a parameter for a generic problem by β and its ML estimate by $\hat{\beta}$. We denote the likelihood function by $\ell(\beta)$. The β value that maximizes $\ell(\beta)$ also maximizes $L(\beta) = \log[\ell(\beta)]$. It is simpler to maximize $L(\beta)$ since it is a sum rather than a product of terms. For many models, $L(\beta)$ has concave shape and $\hat{\beta}$ is the point at which the derivative equals 0. The ML estimate is then the solution of the likelihood equation, $\partial L(\beta)/\partial\beta = 0$. Often, β is multidimensional, denoted by $\boldsymbol{\beta}$, and $\hat{\boldsymbol{\beta}}$ is the solution of a set of likelihood equations.

Let $\text{cov}(\hat{\boldsymbol{\beta}})$ denote the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$. Under regularity conditions (Rao 1973, p. 364), $\text{cov}(\hat{\boldsymbol{\beta}})$ is the inverse of the *information matrix*. The (j, k) element of the information matrix is

$$(1.6) \quad -E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k}\right).$$

The standard errors are the square roots of the diagonal elements for the inverse of the information matrix. The greater the curvature of the log likelihood function, the smaller the standard errors. This is reasonable, since large curvature implies that the log likelihood drops quickly as $\boldsymbol{\beta}$ moves away from $\hat{\boldsymbol{\beta}}$; hence, the data would have been much more likely to occur if $\boldsymbol{\beta}$ took a value near $\hat{\boldsymbol{\beta}}$ rather than a value far from $\hat{\boldsymbol{\beta}}$.

1.3.2 Likelihood Function and ML Estimate for Binomial Parameter

The part of a likelihood function involving the parameters is called the *kernel*. Since the maximization of the likelihood is done with respect to the parameters, the rest is irrelevant.

To illustrate, consider the binomial distribution (1.1). The binomial coefficient $n!/[y!(n-y)!]$ has no influence on where the maximum occurs with respect to π . Thus, we ignore it and treat the kernel as the likelihood function. The binomial log likelihood function is then

$$(1.7) \quad L(\pi) = \log[\pi^y(1-\pi)^{n-y}] = y\log(\pi) + (n-y)\log(1-\pi).$$

Differentiating with respect to π yields

$$(1.8) \quad \partial L(\pi)/\partial\pi = y/\pi - (n-y)/(1-\pi) = (y-n\pi)/\pi(1-\pi).$$

Equating this to 0 gives the likelihood equation, which has solution $\hat{\pi} = y/n$, the sample proportion of successes for the n trials.

Calculating $\partial^2 L(\pi)/\partial\pi^2$, taking the expectation, and combining terms, we get

$$(1.9) \quad -E[\partial^2 L(\pi)/\partial\pi^2] = E[y/\pi^2 + (n-y)/(1-\pi)^2] = n/[\pi(1-\pi)].$$

Thus, the asymptotic variance of $\hat{\pi}$ is $\pi(1-\pi)/n$. This is no surprise. Since $E(Y) = n\pi$ and $\text{var}(Y) = n\pi(1-\pi)$, the distribution of $\hat{\pi} = Y/n$ has mean and standard deviation

$$E(\hat{\pi}) = \pi, \quad \sigma(\hat{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}}.$$

1.3.3 Wald–Likelihood Ratio-Score Test Triad

There are three standard ways to use the likelihood function to perform large-sample inference. We introduce these for a significance test of a null hypothesis $H_0: \beta = \beta_0$ and then discuss their relation to interval estimation. They all exploit the large-sample normality of ML estimators.

Standard errors obtained from the inverse of the information matrix depend on the unknown parameter values. When we substitute the unrestricted ML estimates (i.e., not assuming the null hypothesis) we obtain an estimated standard error of $\hat{\beta}$, which we denote by SE . Denote $-E[\partial^2 L(\beta)/\partial\beta^2]$ (i.e., the information) evaluated at $\hat{\beta}$ by $\iota(\hat{\beta})$. The first large-sample inference method has test statistic using this estimated standard error,

$$z = (\hat{\beta} - \beta_0)/SE, \quad \text{where } SE = 1/\sqrt{\iota(\hat{\beta})}.$$

This statistic has an approximate standard normal distribution when $\beta = \beta_0$. We refer z to the standard normal table to obtain one- or two-sided P -values. Equivalently, for the two-sided alternative, z^2 has an approximate chi-squared null distribution with $df = 1$; the P -value is then the right-tailed chi-squared probability above the observed value. This type of statistic, using the nonnull estimated standard error, is called a *Wald statistic* (Wald 1943).

The multivariate extension² for the Wald test of $H_0: \beta = \beta_0$ has test statistic

$$W = (\hat{\beta} - \beta_0)^T [\text{cov}(\hat{\beta})]^{-1} (\hat{\beta} - \beta_0).$$

The nonnull covariance is based on the curvature (1.6) of the log-likelihood function at $\hat{\beta}$ and typically itself requires estimation. The asymptotic multivariate normal distribution for $\hat{\beta}$ implies an asymptotic chi-squared distribution for W . The df equal the rank of $\text{cov}(\hat{\beta})$, which is the number of nonredundant parameters in β .

A second general-purpose method uses the likelihood function through the ratio of two maximizations: (1) the maximum over the possible parameter values under H_0 , and (2) the maximum over the larger set of parameter values permitting H_0 or an alternative H_a to be true. Let ℓ_0 denote the maximized value of the likelihood function under H_0 , and let ℓ_1 denote the maximized value generally (i.e., under $H_0 \cup H_a$). For instance, for parameters $\beta = (\beta_0, \beta_1)$ and $H_0: \beta_0 = \mathbf{0}$, ℓ_1 is the likelihood function calculated at the β value for which the data would have been most likely; ℓ_0 is the likelihood function calculated at the β_1 value for which the data would have been most likely, when $\beta_0 = \mathbf{0}$. Then ℓ_1 is always at least as large as ℓ_0 , since ℓ_0 results from maximizing over a restricted set of the parameter values.

The ratio $\Lambda = \ell_0/\ell_1$ of the maximized likelihoods cannot exceed 1. Wilks (1935, 1938) showed that $-2 \log \Lambda$ has a limiting null chi-squared distribution, as $n \rightarrow \infty$. The df equal the difference in the dimensions of the parameter spaces under $H_0 \cup H_a$ and under H_0 . The *likelihood-ratio test statistic* equals

$$-2 \log \Lambda = -2 \log(\ell_0/\ell_1) = -2(L_0 - L_1),$$

where L_0 and L_1 denote the maximized log-likelihood functions. [In this book, we use the natural logarithm throughout, for which its inverse is the exponential function; so, if $a = \log(b)$, then $b = \exp(a) = e^a$.]

The third method uses the *score statistic*, due to R. A. Fisher and C. R. Rao. The score test, referred to in some literature as the *Lagrange multiplier test*, is based on the slope and expected curvature of the log-likelihood function $L(\beta)$ at the null value β_0 . It utilizes the size of the *score function*

$$u(\beta) = \partial L(\beta)/\partial\beta,$$

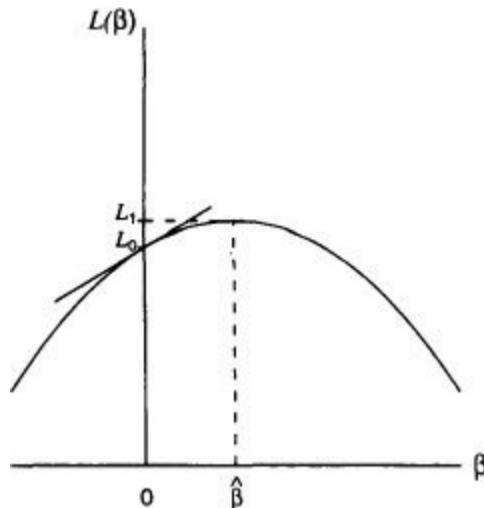
evaluated at β_0 . The value $u(\beta_0)$ tends to be larger in absolute value when $\hat{\beta}$ is farther from β_0 . Denote $-E[\partial^2 L(\beta)/\partial\beta^2]$ evaluated at β_0 by $\iota(\beta_0)$. The score statistic is the ratio of $u(\beta_0)$ to its null SE , which is $[\iota(\beta_0)]^{1/2}$. This has an approximate standard normal null distribution. The chi-squared form of the score statistic is

$$\frac{[u(\beta_0)]^2}{i(\beta_0)} = \frac{[\partial L(\beta)/\partial \beta_0]^2}{-E[\partial^2 L(\beta)/\partial \beta_0^2]},$$

where the notation reflects derivatives with respect to β that are evaluated at β_0 . In the multiparameter case, the score statistic is a quadratic form based on the vector of partial derivatives of the log likelihood with respect to β and the inverse information matrix, both evaluated at the H_0 estimates (i.e., assuming that $\beta = \beta_0$).

[Figure 1.1](#) shows a plot of a generic log-likelihood function $L(\beta)$ for the univariate case. It illustrates the three tests of $H_0: \beta = 0$. The Wald test uses the behavior of $L(\beta)$ at the ML estimate $\hat{\beta}$, having chi-squared form $(\hat{\beta}/SE)^2$. The SE of $\hat{\beta}$ depends on the curvature of $L(\beta)$ at $\hat{\beta}$. The score test is based on the slope and curvature of $L(\beta)$ at $\beta = 0$. The likelihood-ratio test combines information about $L(\beta)$ at both $\hat{\beta}$ and $\beta_0 = 0$. It compares the log-likelihood values L_1 at $\hat{\beta}$ and L_0 at $\beta_0 = 0$ using the chi-squared statistic $-2(L_0 - L_1)$. In [Figure 1.1](#), this statistic is twice the vertical distance between values of $L(\beta)$ at $\hat{\beta}$ and at 0.

[Figure 1.1](#) Log-likelihood function and information used in three tests of $H_0: \beta = 0$.



Section 1.4.1 illustrates the Wald, likelihood-ratio, and score tests for inference about a binomial parameter. As $n \rightarrow \infty$, the three tests have certain asymptotic equivalences (Cox and Hinkley 1974, Sec. 9.3). For small to moderate sample sizes, the likelihood-ratio and score tests are usually more reliable than the Wald test, having actual error rates closer to the nominal level.

1.3.4 Constructing Confidence Intervals by Inverting Tests

In practice, it is more informative to construct confidence intervals for parameters than to test hypotheses about their values. For any of the three test methods, we can construct a confidence interval by inverting the test. For instance, a 95% confidence interval for β is the set of β_0 for which the test of $H_0: \beta = \beta_0$ has P -value exceeding 0.05.

Let z_a denote the z -score from the standard normal distribution having right-tailed probability a ; this is the $100(1 - a)$ percentile of that distribution. A $100(1 - \alpha)\%$ confidence interval based on asymptotic normality uses $z_{\alpha/2}$, for instance $z_{0.025} = 1.96$ for 95% confidence. The Wald confidence interval is the set of β_0 for which $|\hat{\beta} - \beta_0|/SE < z_{\alpha/2}$. This gives the interval $\hat{\beta} \pm z_{\alpha/2}(SE)$. Let $X^2_{df}(a)$ denote the $100(1 - a)$ percentile of the chi-squared distribution with degrees of freedom df . The likelihood-ratio-based confidence interval is the set of β_0 for which $-2[L(\beta_0) - L(\hat{\beta})] < \chi^2_{df}(\alpha)$. [Note that $\chi^2_{df}(\alpha) = z_{\alpha/2}^2$.]

When β has a normal distribution, the log-likelihood function has a parabolic shape. For small samples with categorical data, β may be far from normality and the log-likelihood function can be far from a symmetric, parabolic-shaped curve. This can also happen with moderate to large samples when β falls near the boundary of the parameter space, such as a population proportion that is near 0 or near 1. In such cases, inference based on asymptotic normality of β may have inadequate performance. A marked divergence in results of Wald and likelihood-ratio inference indicates that the distribution of β may not be close to normality. The example in Section 1.4.3 illustrates.

The Wald confidence interval is commonly used in practice, because it is simple to construct using ML estimates and standard errors reported by statistical software. The likelihood-ratio-test-based interval is becoming more widely available in software and is preferable for categorical data with small to moderate n . The score-test-based interval is widely available only in certain cases, such as for proportions as outlined in Section 1.4.2. For the best known statistical model, regression for a normal response, the three types of inference provide identical results. In later chapters, we'll use versions of these intervals that apply for models with multiple parameters. Especially useful is the *profile likelihood* approach based on inverting likelihood-ratio tests (e.g., Section 3.2.6).

1.4 STATISTICAL INFERENCE FOR BINOMIAL PARAMETERS

In this section we illustrate inference methods for categorical data by presenting tests and confidence intervals for the binomial parameter π . With y successes in n independent trials, recall that the ML estimator of π is $\hat{\pi} = y/n$, for which $E(\hat{\pi}) = \pi$ and $\text{var}(\hat{\pi}) = \pi(1 - \pi)/n$.

1.4.1 Tests About a Binomial Parameter

Consider $H_0: \pi = \pi_0$. Since H_0 has a single parameter, we use the normal rather than chi-squared forms of Wald and score test statistics. They permit tests against one-sided as well as two-sided alternatives.

The Wald statistic for testing $H_0: \pi = \pi_0$ is

$$(1.10) \quad z_W = \frac{\hat{\pi} - \pi_0}{SE} = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1 - \hat{\pi})/n}}.$$

To find the score statistic, we evaluate the binomial score (1.8) and information (1.9) at π_0 . This yields

$$u(\pi_0) = \frac{y}{\pi_0} - \frac{n - y}{1 - \pi_0}, \quad i(\pi_0) = \frac{n}{\pi_0(1 - \pi_0)}.$$

The normal form of the score statistic simplifies to

$$(1.11) \quad z_S = \frac{u(\pi_0)}{[i(\pi_0)]^{1/2}} = \frac{y - n\pi_0}{\sqrt{n\pi_0(1 - \pi_0)}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}}.$$

Whereas the Wald statistic z_W uses the standard error evaluated at $\hat{\pi}$, the score statistic z_S uses it evaluated at π_0 . The score statistic is preferable, as it uses the actual null SE rather than an estimate. Its null sampling distribution is closer to standard normal than that of the Wald statistic.

The binomial log-likelihood function (1.7) equals $L_0 = y \log \pi + (n - y) \log(1 - \pi_0)$ under H_0 and $L_1 = y \log \hat{\pi} + (n - y) \log(1 - \hat{\pi})$ more generally. The likelihood-ratio test statistic simplifies to

$$-2(L_0 - L_1) = 2 \left[y \log \frac{\hat{\pi}}{\pi_0} + (n - y) \log \frac{1 - \hat{\pi}}{1 - \pi_0} \right].$$

Expressed as

$$-2(L_0 - L_1) = 2 \left[y \log \frac{y}{n\pi_0} + (n - y) \log \frac{n - y}{n - n\pi_0} \right],$$

it compares observed success and failure counts with fitted counts under H_0 by

$$(1.12) \quad 2 \sum_{\text{observed}} \left[\log \left(\frac{\text{observed}}{\text{fitted}} \right) \right].$$

We'll see that this formula also holds for tests about Poisson and multinomial parameters. Since no unknown parameters occur under H_0 and one occurs under H_a , the asymptotic chi-squared distribution for (1.12) has $df = 1 - 0 = 1$.

1.4.2 Confidence Intervals for a Binomial Parameter

Inverting the Wald test statistic gives the interval of π_0 values for which $|z_W| < z_{\alpha/2}$, or

$$(1.13) \quad \hat{\pi} \pm z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}.$$

Historically, this was one of the first confidence intervals used for any parameter (Laplace 1812, p. 283). Unfortunately, it performs poorly unless n is very large (e.g. Brown et al. 2001), in the sense that the actual probability that the interval contains π usually falls below the nominal confidence coefficient, much below when π is near 0 or 1.

The likelihood-ratio-based confidence interval is more complex computationally, but simple in principle. It is the set of π_0 for which the likelihood-ratio test has a P -value exceeding α . Equivalently, it is the set of π_0 for which double the log likelihood drops by less than $\chi^2_1(\alpha)$ from its value at the ML estimate $\hat{\pi} = y/n$. For example, the endpoints of the 95% confidence interval can be found using numerical methods to iteratively solve for the values of π_0 that satisfy

$$2 \left[y \log \frac{\hat{\pi}}{\pi_0} + (n - y) \log \frac{1 - \hat{\pi}}{1 - \pi_0} \right] = \chi^2_1(0.05) = 3.84.$$

The score confidence interval contains π_0 values for which $|z_S| < z_{\alpha/2}$. Its endpoints are the π_0 solutions to the equations

$$(\hat{\pi} - \pi_0)/\sqrt{\pi_0(1 - \pi_0)/n} = \pm z_{\alpha/2}.$$

These are quadratic in π_0 . First discussed by Wilson (1927), this interval is

$$(1.14) \quad \left[\hat{\pi} \left(\frac{n}{n + z_{\alpha/2}^2} \right) + \frac{1}{2} \left(\frac{z_{\alpha/2}^2}{n + z_{\alpha/2}^2} \right) \right] \pm z_{\alpha/2} \sqrt{\frac{1}{n + z_{\alpha/2}^2} \left[\hat{\pi}(1 - \hat{\pi}) \left(\frac{n}{n + z_{\alpha/2}^2} \right) + \left(\frac{1}{2} \right) \left(\frac{1}{2} \right) \left(\frac{z_{\alpha/2}^2}{n + z_{\alpha/2}^2} \right) \right]}.$$

The midpoint is a weighted average of $\hat{\pi}$ and $\frac{1}{2}$, where the weight $n/(n + z_{\alpha/2}^2)$ given $\hat{\pi}$ increases as n increases. Combining terms, this midpoint equals $\hat{\pi} = (y + z_{\alpha/2}^2/2)/(n + z_{\alpha/2}^2)$. This is the sample proportion for an adjusted sample that adds $z_{\alpha/2}^2$ observations, half of each type, for example, $z_{0.025}^2/2 = 1.96^2/2 \approx 2$ of each type for 95% intervals. The square of the coefficient of $z_{\alpha/2}$ in (1.14) is a weighted average of the variance of a sample proportion when $\pi = \hat{\pi}$ and the variance of a sample proportion when $\pi = \frac{1}{2}$, using the adjusted sample size $n + z_{\alpha/2}^2$ in place of n .

For 95% confidence, the score interval can be approximated by a simple adjustment of the Wald interval (see Exercise 1.25) that adds 2 observations of each type to the sample before using the Wald formula (1.13). This interval and the ordinary score interval tend to have actual coverage probability much closer to the nominal level than the Wald interval (Agresti and Coull 1998, Agresti and Caffo 2000).

1.4.3 Example: Estimating the Proportion of Vegetarians

To collect data to illustrate concepts in introductory statistics courses, often I have given the students a questionnaire. One year I asked each student in an honors class at the University of Florida whether he or she was a vegetarian. Of $n = 25$ students, $y = 0$ answered “yes.” They were not a random sample of a particular population, but we use these data to illustrate 95% confidence intervals for a binomial parameter π .

Since $y = 0$, the ML estimate $\hat{\pi} = 0/25 = 0$. With the Wald method, the 95% confidence interval for π is

$$\hat{\pi} \pm 1.96\sqrt{\hat{\pi}(1 - \hat{\pi})/n}, \text{ which is } 0 \pm 1.96\sqrt{(0.0 \times 1.0)/25}, \text{ or } (0, 0).$$

When a parameter falls near the boundary of the sample space, often sample estimates of standard errors are poor and the Wald method does not provide a sensible answer.

By contrast, the 95% score interval equals $(0.0, 0.133)$. That is, when $\hat{\pi} = 0.0$ and $n = 25$, the two roots for π_0 that satisfy the equation

$$|\hat{\pi} - \pi_0| = 1.96\sqrt{\pi_0(1 - \pi_0)/n}$$

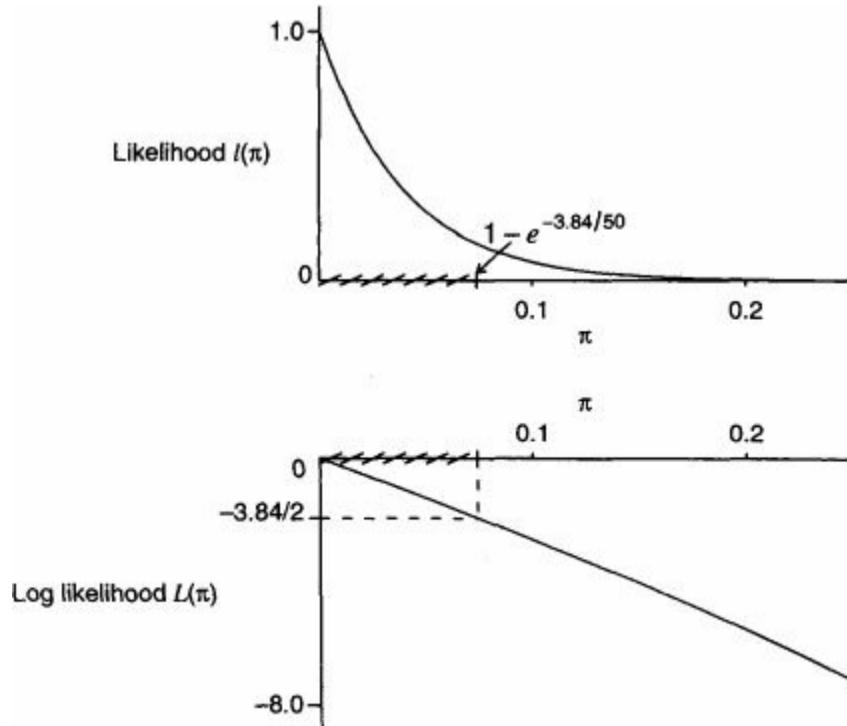
are $\pi_0 = 0.0$ and $\pi_0 = 0.133$. This interval provides a more believable inference. It contains the values not rejected in corresponding score tests with size (probability of type I error) 0.05. For $H_0: \pi = 0.20$, for instance, the score test statistic is $z_S = (0 - 0.20)/\sqrt{(0.20 \times 0.80)/25} = -2.50$, which has two-sided P -value $0.012 < 0.05$, so 0.20 does not fall in the interval. By contrast, for $H_0: \pi = 0.10$, $z_S = (0 - 0.10)/\sqrt{(0.10 \times 0.90)/25} = -1.67$ which has P -value $0.096 > 0.05$, so 0.10 falls in the interval.

When $y = 0$ and $n = 25$, the kernel of the likelihood function is $\ell(\pi) = \pi^0(1 - \pi)^{25} = (1 - \pi)^{25}$. The log-likelihood function (1.7) is $L(\pi) = 25 \log(1 - \pi)$. Note that $L(\hat{\pi}) = L(0) = 0$. The 95% likelihood-ratio confidence interval is the set of π_0 for which the likelihood-ratio statistic

$$\begin{aligned} -2(L_0 - L_1) &= -2[L(\pi_0) - L(\hat{\pi})] \\ &= -50 \log(1 - \pi_0) < \chi^2_{1}(0.05) = 3.84. \end{aligned}$$

The upper bound is $1 - \exp(-3.84/50) = 0.074$, and the confidence interval equals $(0.0, 0.074)$. Figure 1.2 shows the likelihood and log-likelihood functions and the corresponding confidence region for π .

Figure 1.2 Binomial likelihood and log likelihood when $y = 0$ in $n = 25$ trials, and likelihood-ratio test-based confidence interval for π .



The three large-sample methods yield quite different results. When π is near 0, the sampling distribution of $\hat{\pi}$ is highly skewed to the right for small n . From numerical evaluations, we prefer the

interval based on inverting the score test.

1.4.4 Exact Small-Sample Inference and the Mid *P*-Value

With modern computational power, it is not necessary to rely on large-sample approximations for the distribution of estimators such as $\hat{\pi}$. Tests and confidence intervals can directly use the binomial distribution rather than its normal approximation. Such inferences occur naturally for small samples, but apply for any n .

We illustrate by testing $H_0: \pi = 0.50$ against $H_a: \pi \neq 0.50$ for the survey results on vegetarianism just discussed, namely, $y = 0$ with $n = 25$. We noted that the score statistic equals $z = -5.0$. The exact *P*-value for this statistic, based on the null $\text{bin}(25, 0.50)$ distribution, is

$$P(|z| \geq 5.0) = P(Y = 0 \text{ or } Y = 25) = 0.50^{25} + 0.50^{25} = 0.00000006.$$

Because of discreteness, in testing $H_0: \pi = \pi_0$, it is not usually possible to achieve a particular fixed size such as 0.05. With a finite number of possible samples, there is a finite number of possible *P*-values, of which 0.05 may not be one. When $n = 25$ and $\pi_0 = 0.50$, for example, the two-sided *P*-value using the binomial probabilities is 0.043 if $y = 7$ or if $y = 18$ and it is 0.108 if $y = 8$ or if $y = 17$. Thus, if we reject H_0 when $y \leq 7$ or $y \geq 18$, the test is *conservative*, in the sense that the actual size (i.e., 0.043) is less than the nominal size (0.05).

To adjust somewhat for discreteness in small-sample distributions, we can base inference on the *mid P-value* (Lancaster 1949b, 1961). For a test statistic T with observed value t_0 and one-sided H_a such that large T contradicts H_0 ,

$$\text{mid } P\text{-value} = \frac{1}{2}P(T = t_0) + P(T > t_0),$$

with probabilities calculated from the null distribution. Thus, the mid *P*-value is less than the ordinary *P*-value by half the probability of the observed result. Although discrete, compared with the ordinary *P*-value, the mid *P*-value behaves more like the *P*-value for a test statistic having a continuous distribution: The sum of its two one-sided *P*-values equals 1.0. Under H_0 , it has a null expected value of 0.50 (like the uniform distribution that occurs in the continuous case), whereas this expected value exceeds 0.50 for the ordinary *P*-value for a discrete test statistic.

Unlike an exact test with ordinary *P*-value, a test using the mid *P*-value does not guarantee that the size of the test is no greater than a nominal value (Exercise 1.12). However, it usually performs well. It is less conservative than the ordinary exact test. Inference based on the mid *P*-value compromises between the conservativeness of exact methods and the uncertain adequacy of large-sample methods.

Similarly, we can use small-sample distributions to construct confidence intervals for parameters. Some subtle issues arise such that the choice of such an interval is not straightforward, and we defer this topic to a special section (16.6) in Chapter 16 about small-sample intervals for categorical data.

1.5 STATISTICAL INFERENCE FOR MULTINOMIAL PARAMETERS

Next we consider inference for multinomial parameters $\{\pi_j\}$. Of n observations in c categories, n_j occur in category $j, j = 1, \dots, c$.

1.5.1 Estimation of Multinomial Parameters

First, we obtain ML estimates of $\{\pi_j\}$. As a function of $\{\pi_j\}$, the multinomial probability mass function (1.2) is proportional to the kernel

$$(1.15) \quad \prod_j \pi_j^{n_j}, \quad \text{where all } \pi_j \geq 0 \quad \text{and} \quad \sum_j \pi_j = 1.$$

The ML estimates are the $\{\pi_j\}$ that maximize (1.15).

The multinomial log-likelihood function is

$$L(\boldsymbol{\pi}) = \sum_j n_j \log \pi_j.$$

To eliminate redundancies, we treat L as a function of $(\pi_1, \dots, \pi_{c-1})$, since $\pi_c = 1 - (\pi_1 + \dots + \pi_{c-1})$. Thus, $\partial \pi_c / \partial \pi_j = -1$, $j = 1, \dots, c-1$. Since

$$\frac{\partial \log \pi_c}{\partial \pi_j} = \frac{1}{\pi_c} \frac{\partial \pi_c}{\partial \pi_j} = -\frac{1}{\pi_c},$$

differentiating $L(\boldsymbol{\pi})$ with respect to π_j gives the likelihood equation

$$\frac{\partial L(\boldsymbol{\pi})}{\partial \pi_j} = \frac{n_j}{\pi_j} - \frac{n_c}{\pi_c} = 0.$$

The ML solution satisfies $\hat{\pi}_j / \hat{\pi}_c = n_j / n_c$. Now

$$\sum_j \hat{\pi}_j = 1 = \frac{\hat{\pi}_c (\sum_j n_j)}{n_c} = \frac{\hat{\pi}_c n}{n_c},$$

so $\hat{\pi}_c = n_c / n$ and then $\hat{\pi}_j = n_j / n$. From general results presented later in the book (Section 9.6), this solution does maximize the likelihood. Thus, the ML estimates of $\{\pi_j\}$ are the sample proportions.

1.5.2 Pearson Chi-Squared Test of a Specified Multinomial

In 1900 the eminent British statistician Karl Pearson introduced a hypothesis test that was one of the first inferential methods. It had a revolutionary impact on categorical data analysis. Pearson's test evaluates whether multinomial parameters equal certain values. His original motivation in developing this test was to analyze whether possible outcomes on a particular Monte Carlo roulette wheel were equally likely (Stigler 1986).

Consider $H_0: \pi_j = \pi_{j0}, j = 1, \dots, c$, where $\sum_j \pi_{j0} = 1$. When H_0 is true, the expected values of $\{n_j\}$, called *expected frequencies*, are $\mu_j = n\pi_{j0}, j = 1, \dots, c$. Pearson proposed the test statistic

$$(1.16) \quad X^2 = \sum_j \frac{(n_j - \mu_j)^2}{\mu_j}.$$

Greater differences $|n_j - \mu_j|$ produce greater X^2 values, for fixed $\{\pi_{j0}\}$ and n . Let X^2_o denote the observed value of X^2 . The P -value is the null value of $P(X^2 \geq X^2_o)$. This equals the sum of the null multinomial probabilities of all count arrays (having a sum of n) with $X^2 \geq X^2_o$.

For large samples, X^2 has approximately a chi-squared distribution with $df = c - 1$. The P -value is approximated by $P(\chi^2_{c-1} \geq X^2_o)$, where χ^2_{c-1} denotes a chi-squared random variable with $df = c - 1$. Statistic (1.16) is called the *Pearson chi-squared statistic*.

1.5.3 Likelihood-Ratio Chi-Squared Test of a Specified Multinomial

An alternative test for multinomial parameters uses the likelihood-ratio test. The kernel of the multinomial likelihood is (1.15). Under H_0 the likelihood is maximized when $\hat{\pi}_j = \pi_{j0}$. In the general case, it is maximized when $\hat{\pi}_j = n_j/n$. The ratio of the likelihoods equals

$$\Lambda = \frac{\prod_j (\pi_{j0})^{n_j}}{\prod_j (n_j/n)^{n_j}}.$$

Thus, the likelihood-ratio statistic, denoted by G^2 , is

$$(1.17) \quad G^2 = -2 \log \Lambda = 2 \sum_j n_j \log(n_j/n\pi_{j0}).$$

This statistic, which has form (1.12), is called the *likelihood-ratio chi-squared statistic*. The larger the value of G^2 , the greater the evidence against H_0 .

In the general case, the parameter space consists of $\{\pi_j\}$ subject to $\sum_j \pi_j = 1$, so the dimensionality is $c - 1$. Under H_0 , the $\{\pi_j\}$ are specified completely, so the dimension is 0. The difference in these dimensions equals $(c - 1)$. For large n , G^2 has a chi-squared null distribution with $\text{df} = c - 1$.

When H_0 holds, the Pearson X^2 and the likelihood ratio G^2 both have large-sample chi-squared distributions with $\text{df} = c - 1$. In fact, they are asymptotically equivalent in that case; specifically, $X^2 - G^2$ converges in probability to zero. [This means that for any $\epsilon > 0$, $P(|X^2 - G^2| < \epsilon) \rightarrow 1$ as $n \rightarrow \infty$. See Section 16.3.4.] When H_0 is false, X^2 and G^2 grow in expectation proportionally to n ; they need not take similar values, however, even for very large n .

For fixed c , as n increases the distribution of X^2 usually converges to chi-squared more quickly than that of G^2 . The chi-squared approximation is often poor for G^2 when $n/c < 5$. When c is large, it can be decent for X^2 for n/c as small as 1 if the table does not contain both very small and moderately large expected frequencies.

Alternatively, the multinomial probabilities induce exact distributions of these test statistics. When it is not feasible to quickly enumerate all the possible samples, it is simple to simulate the exact distributions by randomly generating a very large number of multinomial samples of size n with the null probabilities, and calculating X^2 and or G^2 for each sample (Hirji 2005, Chap. 13). The simulated P -value is the proportion of test statistic values that are at least as large as the observed value.

1.5.4 Example: Testing Mendel's Theories

Among its many applications, Pearson's test was used in genetics to test Mendel's theories of natural inheritance. Mendel crossed pea plants of pure yellow strain with plants of pure green strain. He predicted that second-generation hybrid seeds would be 75% yellow and 25% green, yellow being the dominant strain. One experiment produced $n = 8023$ seeds, of which $n_1 = 6022$ were yellow and $n_2 = 2001$ were green. The expected frequencies for H_0 : $\pi_{10} = 0.75$, $\pi_{20} = 0.25$ are $\mu_1 = 8023(0.75) = 6017.25$ and $\mu_2 = 2005.75$. The Pearson statistic $X^2 = 0.015$ and the likelihood-ratio statistic $G^2 = 0.015$ ($df = 1$) have P -values of $P = 0.90$. They do not contradict Mendel's hypothesis.

When $c = 2$, Pearson's X^2 simplifies to the square of the normal score statistic (1.11). For Mendel's data, $\hat{\pi}_1 = 6022/8023$, $\pi_{10} = 0.75$, $n = 8023$, and $z_S = 0.123$, for which $X^2 = (0.123)^2 = 0.015$. In fact, for general c the Pearson test is the score test about specified values for multinomial parameters.

Mendel performed several experiments of this type. In 1936, R. A. Fisher summarized Mendel's results. He used the reproductive property of chi-squared: If X^2_1, \dots, X^2_k are independent chi-squared statistics with degrees of freedom v_1, \dots, v_k , then $\sum_{i=1}^k X_i^2$ has a chi-squared distribution with $df = \sum_{i=1}^k v_i$. Fisher obtained a summary chi-squared statistic equal to 42, with $df = 84$. A chi-squared distribution with $df = 84$ has mean 84 and standard deviation $(2 \times 84)^{1/2} = 13.0$, and the right-tailed probability above 42 is $P = 0.99996$. In other words, the chi-squared statistic was so small that the fit seemed *too* good.

Fisher commented: "The general level of agreement between Mendel's expectations and his reported results shows that it is closer than would be expected in the best of several thousand repetitions I have no doubt that Mendel was deceived by a gardening assistant, who knew only too well what his principal expected from each trial made." In a letter written at the time, he stated: "Now, when data have been faked, I know very well how generally people underestimate the frequency of wide chance deviations, so that the tendency is always to make them agree too well with expectations" (Box 1978, p. 297). In summary, goodness-of-fit tests can reveal not only when a fit is inadequate, but also when it is better than random fluctuations would have us expect. [Fisher's daughter, Joan Fisher Box (1978, pp. 295–300), discussed Fisher's analysis of Mendel's data and the accompanying controversy. See also Pires and Branco (2010). Despite possible difficulties with Mendel's data, subsequent work led to general acceptance of his theories.]

1.5.5 Testing with Estimated Expected Frequencies

The chi-squared statistics (1.16) and (1.17) compare a sample distribution to a hypothetical one $\{\pi_{j0}\}$. In some applications, $\{\pi_{j0} = \pi_{j0}(\theta)\}$ are functions of a smaller set of unknown parameters θ . ML estimates $\hat{\theta}$ of θ determine ML estimates $\{\pi_{j0}(\hat{\theta})\}$ of $\{\pi_{j0}\}$ and hence ML estimates $\{\hat{\mu}_j = n\pi_{j0}(\hat{\theta})\}$ of expected frequencies.

Replacing $\{\mu_j\}$ by estimates $\{\hat{\mu}_j\}$ affects the distribution of X^2 and G^2 . When $\dim(\theta) = p$, the true df = $(c - 1) - p$ (Section 16.3.3). Pearson (1917) realized this but did not always take it into account (Section 17.2).

1.5.6 Example: Pneumonia Infections in Calves

We now show a goodness-to-fit test with estimated expected frequencies. A sample of 156 dairy calves born in Okeechobee County, Florida, were classified according to whether they caught pneumonia within 60 days of birth. Calves that got a pneumonia infection were also classified according to whether they got a secondary infection within 2 weeks after the first infection cleared up. [Table 1.1](#) shows the data. Calves that did not get a primary infection could not get a secondary infection, so no observations can fall in the category for “no” primary infection and “yes” secondary infection. That combination is called a *structural zero*.

Table 1.1 Primary and Secondary Pneumonia Infections in Calves

		Secondary Infection ^a	
		Yes	No
Primary Infection	Yes	30 (38.1)	63 (39.0)
	No	0 (—)	63 (78.9)

^aValues in parentheses are estimated expected frequencies.

Source: Data courtesy of Thang Tran and G. A. Donovan, College of Veterinary Medicine, University of Florida.

A goal of this study was to test whether the probability of primary infection was the same as the conditional probability of secondary infection, given that the calf got the primary infection. In other words, if π_{ab} denotes the probability that a calf is classified in row a and column b of this table, the null hypothesis is

$$H_0: \pi_{11} + \pi_{12} = \pi_{11}/(\pi_{11} + \pi_{12})$$

or $\pi_{11} = (\pi_{11} + \pi_{12})^2$. Let $\pi = \pi_{11} + \pi_{12}$ denote the probability of primary infection. The null hypothesis states that the probabilities satisfy the structure that [Table 1.2](#) shows; that is, probabilities in a trinomial for the categories (yes–yes, yes–no, no–no) for primary–secondary infection equal $[\pi^2, \pi(1 - \pi), 1 - \pi]$.

Table 1.2 Probability Structure for Hypothesis

		Secondary Infection		Total
		Yes	No	
Primary Infection	Yes	π^2	$\pi(1 - \pi)$	π
	No	—	$1 - \pi$	$1 - \pi$

Let n_{ab} denote the number of observations in row a and column b of [Table 1.1](#). The ML estimate of π is the value maximizing the kernel of the multinomial likelihood

$$(\pi^2)^{n_{11}}(\pi - \pi^2)^{n_{12}}(1 - \pi)^{n_{22}}.$$

The log likelihood is

$$L(\pi) = n_{11} \log \pi^2 + n_{12} \log(\pi - \pi^2) + n_{22} \log(1 - \pi).$$

Differentiation with respect to π gives the likelihood equation

$$\frac{2n_{11}}{\pi} + \frac{n_{12}}{\pi} - \frac{n_{12}}{1 - \pi} - \frac{n_{22}}{1 - \pi} = 0.$$

The solution is

$$\hat{\pi} = (2n_{11} + n_{12})/(2n_{11} + 2n_{12} + n_{22}).$$

For [Table 1.1](#), $\hat{\pi} = 0.494$. Since $n = 156$, the estimated expected frequencies are $\hat{n}_{11} = n\hat{\pi}^2 = 38.1$, $\hat{n}_{12} = n(\hat{\pi} - \hat{\pi}^2) = 39.0$, and $\hat{n}_{22} = n(1 - \hat{\pi}) = 78.9$. [Table 1.1](#) shows them. Pearson's statistic is $X^2 = 19.7$. Since the $c = 3$ possible responses have $p = 1$ parameter (π) determining the expected frequencies, $df = (3 - 1) - 1 = 1$. There is strong evidence against H_0 ($P = 0.00001$). Inspection of [Table 1.1](#) reveals that many more calves got a primary infection but not a secondary infection than H_0 predicts. The researchers concluded that the primary infection had an immunizing effect that reduced the likelihood of a secondary infection.

1.5.7 Chi-Squared Theoretical Justification

We now outline why Pearson's statistic for a specified multinomial has a limiting chi-squared distribution. Derivations for the likelihood-ratio statistic and cases with estimated expected frequencies are given in Section 16.3.

For a multinomial sample (n_1, \dots, n_c) of size n , the marginal distribution of n_j is the $\text{bin}(n, \pi_j)$ distribution. For large n , by the normal approximation to the binomial, n_j (and $\hat{\pi}_j = n_j/n$) have approximate normal distributions. More generally, by the central limit theorem, the sample proportions $\hat{\pi} = (n_1/n, \dots, n_{c-1}/n)^T$ have an approximate multivariate normal distribution (Section 16.1.4). Let Σ_0 denote the null covariance matrix of $\sqrt{n}\hat{\pi}$, and let $\pi_0 = (\pi_{10}, \dots, \pi_{c-1,0})^T$. Under H_0 , since $\sqrt{n}(\hat{\pi} - \pi_0)$ converges to a $N(\mathbf{0}, \Sigma_0)$ distribution, the quadratic form

$$(1.18) \quad n(\hat{\pi} - \pi_0)^T \Sigma_0^{-1}(\hat{\pi} - \pi_0)$$

has distribution converging to chi-squared with $\text{df} = c - 1$.

In Section 16.1.4 we show that the covariance matrix of $\sqrt{n}\hat{\pi}$ has elements

$$\sigma_{jk} = \begin{cases} -\pi_j \pi_k & \text{if } j \neq k \\ \pi_j(1 - \pi_j) & \text{if } j = k \end{cases}$$

The matrix Σ_0^{-1} has (j, k) th element $1/\pi_{c0}$ when $j \neq k$ and $(1/\pi_{j0} + 1/\pi_{c0})$ when $j = k$. (You can verify this by showing that $\Sigma_0 \Sigma_0^{-1}$ equals the identity matrix.) With this substitution, direct calculation with appropriate combining of terms yields that (1.18) simplifies to X^2 . In Section 16.3 we provide a formal proof in a more general setting.

This argument is similar to Pearson's in 1900. R. A. Fisher (1922) gave a simpler justification, the gist of which follows: Suppose that (n_1, \dots, n_c) are independent Poisson random variables with means (μ_1, \dots, μ_c) . For large $\{\mu_j\}$, the standardized values $\{z_j = (n_j - \mu_j)/\sqrt{\mu_j}\}$ have approximate standard normal distributions. Thus, $\sum_j z_j^2 = X^2$ has an approximate chi-squared distribution with c degrees of freedom. Adding the single linear constraint $\sum_j (n_j - \mu_j) = 0$ thus converting the Poisson distributions to a multinomial, we lose a degree of freedom.

1.6 BAYESIAN INFERENCE FOR BINOMIAL AND MULTINOMIAL PARAMETERS

This book mainly uses the traditional, so-called *frequentist*, approach to statistical inference. We regard parameter values as fixed and apply probability statements to possible values for the data, given the parameter values. Recent years have seen increasing popularity of the *Bayesian* approach, which has probability distributions for parameters as well as for data. This yields inferences in the form of probability statements about possible values for the parameters, given the data.

1.6.1 The Bayesian Approach to Statistical Inference

The Bayesian approach assumes a *prior distribution* for the parameters. This probability distribution may reflect subjective prior beliefs. Or, it may reflect information about the parameter values from other studies. Or, it may be relatively uninformative, so that inferential results are based almost entirely on the current data. The prior distribution combines with the information that the data provide to generate a *posterior distribution* for the parameters. Different choices for the prior distribution can result in quite different posterior inferences, especially for small sample sizes, so the choice should be given careful thought.

By Bayes' theorem, the posterior probability density function h of a parameter θ , given the data y , relates to the probability mass function f for y , given θ , and the prior density function g for θ , by

$$h(\theta | y) = \frac{f(y | \theta)g(\theta)}{f(y)}.$$

The denominator $f(y)$ on the right-hand side is the marginal probability mass function of the data, that is, $\int_{\Theta} f(y | \theta)g(\theta)d\theta$. This is a constant with respect to θ , so irrelevant for inference about θ . When we plug in the observed data, $f(y | \theta)$ is the likelihood function when viewed as a function of θ . So, the prior density function for θ multiplied by the likelihood function determines the posterior density for θ .

Except in specialized cases such as presented in Sections 1.6.2 and 1.6.3, there is not a closed-form expression for the posterior distribution. The difficulty is in finding the denominator integral that determines $f(y)$. The key part of the Bayes equation is the numerator, because of the proportionality in terms of θ ,

$$h(\theta | y) \propto f(y | \theta)g(\theta).$$

Simulation methods are used to approximate the posterior distribution. The primary method for doing this is Markov chain Monte Carlo (MCMC). It is beyond our scope to discuss the technical details of how an MCMC algorithm works. In a nutshell, a stochastic process of Markov chain form is designed so that its long-run stationary distribution is the posterior distribution. One or more such Markov chains provide a very large number of simulated values from the posterior distribution, and the distribution of the simulated values approximates the posterior distribution. Enough observations are taken after a burn-in period so that the Monte Carlo error is small in approximating the posterior distribution and summary measures of interest for that distribution, such as the mean and standard deviation, certain percentiles, and intervals formed using those percentiles.

For an arbitrary parameter β , such as a coefficient in a regression-type model, Bayesian methods of inference using the posterior distribution parallel those for frequentist inference. For example, in lieu of P -values, posterior tail probabilities are useful. Information about the direction of an effect is contained in the posterior probabilities $P(\beta > 0 | y)$ and $P(\beta < 0 | y)$. With a flat prior distribution, $P(\beta < 0 | y)$ corresponds to the frequentist P -value for the one-sided test with $H_a: \beta > 0$.

Analogous to the frequentist confidence interval is an interval that contains most of the posterior distribution. Such an interval is referred to as a *posterior interval* or *credible interval*. A common approach for constructing a posterior interval uses percentiles of the posterior distribution, with equal probabilities in the two tails. For example, the 95% equal-tail posterior interval for β is the region between the 2.5 and 97.5 percentiles of the posterior distribution for β . For unimodal posteriors, an alternative Bayesian *highest posterior density* (HPD) interval has higher posterior density for every value inside the interval than for every value outside it, subject to the posterior probability over the interval equaling the desired confidence level. This method produces the shortest possible interval with the given level.

We next summarize the Bayesian approach for binomial and multinomial parameters. Then, in the rest of the book, we'll occasionally present Bayesian alternatives to frequentist model-based inference.

1.6.2 Binomial Estimation: Beta and Logit-Normal Prior Distributions

The simplest Bayesian inference for a binomial parameter π uses a member of the *beta distribution* as the prior distribution. The $\text{beta}(\alpha_1, \alpha_2)$ probability density function for π is proportional to

$$\pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}.$$

The parameters $\alpha_1 > 0$ and $\alpha_2 > 0$ of the prior are often referred to as *hyperparameters*, to distinguish them from the parameter that is the object of inference (in this case, π). The beta distribution has

$$E(\pi) = \alpha_1/(\alpha_1 + \alpha_2) \quad \text{and} \quad \text{var}(\pi) = \alpha_1\alpha_2/[(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)].$$

The family of beta probability density functions has a wide variety of shapes over the interval $(0, 1)$, including uniform when $\alpha_1 = \alpha_2 = 1$, unimodal symmetric ($\alpha_1 = \alpha_2 > 1$), unimodal skewed left ($\alpha_1 > \alpha_2 > 1$), unimodal skewed right ($\alpha_2 > \alpha_1 > 1$), and bimodal U-shaped ($\alpha_1 < 1, \alpha_2 < 1$).

Often prior knowledge about π can be expressed in terms of a mean and standard deviation for a prior for π . Then, the one-to-one correspondence between those moments and (α_1, α_2) based on the above moment expressions determines a beta prior. By contrast, lack of prior knowledge about π might suggest using a uniform prior distribution. The posterior distribution then has the same shape as the binomial likelihood function. Alternatively, a popular prior distribution with Bayesians is the *Jeffreys prior*, which is proportional to the square root of the determinant of the Fisher information matrix for the parameters of interest. With this approach, prior distributions for different scales of measurement for the parameters (e.g., for π or for $\phi = \log[\pi/(1 - \pi)]$) are equivalent. For a binomial parameter, the Jeffreys prior is the beta distribution with $\alpha_1 = \alpha_2 = 0.5$.

The beta distribution is the *conjugate prior distribution* for inference about a binomial parameter. This means that it is the family of probability distributions such that, when combined with the likelihood function, the posterior distribution falls in the same family. When we combine a $\text{beta}(\alpha_1, \alpha_2)$ prior distribution with a binomial likelihood function, the posterior distribution is a $\text{beta}(y + \alpha_1, n - y + \alpha_2)$ distribution, for which the mean is

$$\frac{y + \alpha_1}{n + \alpha_1 + \alpha_2} = \left(\frac{n}{n + \alpha_1 + \alpha_2} \right) \hat{\pi} + \left(\frac{\alpha_1 + \alpha_2}{n + \alpha_1 + \alpha_2} \right) \frac{\alpha_1}{\alpha_1 + \alpha_2}.$$

This is a weighted average of the sample proportion $\hat{\pi} = y/n$ and the prior mean, with more weight given the sample proportion as n increases. Conjugate priors were the primary method of conducting Bayesian analysis before the development of computationally intensive methods, such as Markov chain Monte Carlo, for evaluating the integral that determines the posterior distribution.

An alternative prior distribution assumes a normal distribution for the *logit* parameter, $\log[\pi/(1 - \pi)]$. This parameter, which is relevant for many analyses presented in this book, takes values over the entire real line. With a $N(0, \sigma^2)$ prior distribution for $\log[\pi/(1 - \pi)]$, on the π scale the shape of this *logit-normal* (also called *logistic-normal*) density is symmetric³, being unimodal when $\sigma^2 \leq 2$ and bimodal when $\sigma^2 > 2$, but always tapering off toward 0 as π approaches 0 or 1. Specifically, it is mound-shaped for small σ , roughly uniform except near the boundaries when $\sigma \approx 1.5$, and with more pronounced peaks for the modes when σ is about 2 or larger. The peaks for the modes get closer to 0 and 1 as σ increases further, and the curve has appearance that is essentially U-shaped when $\sigma = 3$ and similar to that of a $\text{beta}(0.5, 0.5)$ prior. For $\sigma = (1, 2, 3)$, the standard deviations on the π scale of these priors are $(0.21, 0.31, 0.37)$, similar to the values $(0.22, 0.29, 0.35)$ for the beta priors with $\alpha_1 = \alpha_2 = (2.0, 1.0, 0.5)$. The logit-normal prior with $\sigma = 2.67$ matches the Jeffreys prior in the first two moments (on the probability scale), and the logit-normal prior with $\sigma = 1.69$ matches the uniform prior in the first two moments. With a $N(\mu, \sigma^2)$ prior distribution for the logit, the density for π is skewed left when $\mu > 0$ and skewed right when $\mu < 0$.

Yet another possibility, hierarchical in nature, uses beta or logit-normal priors but assumes a distribution for their hyperparameters instead of assigning fixing values. That second-stage distribution may have its own hyperparameters. See Section 3.6.7, Albert (2010), Good (1965), and

Leonard (1972).

1.6.3 Multinomial Estimation: Dirichlet Prior Distributions

For $c > 2$ categories, the beta distribution generalizes to the *Dirichlet distribution*. It is defined over the simplex of nonnegative values $\boldsymbol{\pi} = (\pi_1, \dots, \pi_c)$ that sum to 1. Expressed in terms of gamma functions and c hyperparameters $\{\alpha_i > 0\}$, the Dirichlet probability density function is

$$g(\boldsymbol{\pi}) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^c \pi_i^{\alpha_i - 1} \quad \text{for } 0 < \pi_i < 1 \text{ all } i, \quad \sum_i \pi_i = 1.$$

The case $\{\alpha_i = 1\}$ is the uniform density over the possible probability values. The case $\{\alpha_i = \frac{1}{2}\}$ is the *Jeffreys prior* for multinomial parameters. Let $K = \sum_i \alpha_i$. The Dirichlet distribution has $E(\pi_i) = \alpha_i/K$ and $\text{var}(\pi_i) = \alpha_i(K - \alpha_i)/[K^2(K + 1)]$. For particular relative sizes of $\{\alpha_i\}$, such as identical values, the distribution is more tightly concentrated around the means as K increases.

Let $\mathbf{n} = (n_1, \dots, n_c)$ denote cell counts from $n = \sum_i n_i$ independent observations with cell probabilities $\boldsymbol{\pi}$. Formula (1.2) showed the multinomial probability mass function for \mathbf{n} . Multiplying this by the Dirichlet prior density function $g(\boldsymbol{\pi})$ contributes to a posterior density function $h(\boldsymbol{\pi} | \mathbf{n})$ for $\boldsymbol{\pi}$ that is also Dirichlet, but with the hyperparameters $\{\alpha_i\}$ replaced by $\{\alpha'_i = n_i + \alpha_i\}$. The mean of the posterior distribution of π_i is

$$E(\pi_i | n_1, \dots, n_c) = (n_i + \alpha_i)/(n + K).$$

Let $\gamma_i = E(\pi_i) = \alpha_i/K$. This Bayesian estimator equals the weighted average

$$(1.19) \quad \left(\frac{n}{n+K} \right) p_i + \left(\frac{K}{n+K} \right) \gamma_i$$

of the sample proportion $p_i = n_i/n$ and the mean γ_i of the prior distribution for π_i . This posterior mean takes the form of a sample proportion when the prior information corresponds to K additional observations of which α_i were outcomes of type i . (We'll consider a formal way of setting such *data augmentation priors* in Section 7.2.4.) With identical $\{\alpha_i\}$, the Bayes estimate shrinks each sample proportion toward the equi-probability value $\gamma_i = 1/c$. Greater shrinkage occurs as K increases, for fixed n .

Bayesian estimators of multinomial parameters, unlike the sample proportions, are slightly biased for finite n . Usually, though, they have smaller total mean squared error (MSE) than the sample proportions. They are not uniformly better for all possible parameter values, however. For instance, if a particular $\pi_i = 0$, then $p_i = 0$ with probability one, so the sample proportion is then better than any other estimator. We do not expect $\pi_i = 0$ in practice, and the parameter space is often defined under the restriction that all $\pi_i > 0$, but this limiting behavior explains why the ML estimator can have smaller MSE than the Bayes estimator when π_i is very near 0.

1.6.4 Example: Estimating Vegetarianism Revisited

In Section 1.4.3 we estimated the population proportion of vegetarians with a sample of size $n = 25$ for which $y = 0$. The ML estimate of π is $\hat{\pi} = 0.0$, and the 95% score confidence interval is $(0.0, 0.133)$. How does this compare to Bayesian point and interval estimates?

First, we use a uniform prior distribution for π , reflecting prior ignorance. For this $\text{beta}(1, 1)$ prior with $y = 0$ and $n = 25$, the posterior distribution is $\text{beta}(1, 26)$. The posterior mean is $1/27 = 0.037$. The posterior 95% equal-tail interval is $(0.001, 0.132)$, the endpoints being the 2.5 and 97.5 percentiles of the beta posterior density. This interval is similar to the frequentist 95% score interval, but the prior information has the impact of moving the left boundary slightly away from 0.0. By contrast, since the posterior density is proportional to $(1 - \pi)^{25}$ and hence monotone decreasing, the 95% highest posterior density (HPD) interval has lower limit of 0 and upper limit that is the 95th percentile of the $\text{beta}(1, 26)$ density, which is 0.109.

For contrast, let's use a much more informative beta prior. Suppose we used a subjective approach and were quite sure *a priori* that π falls between about 0 and 0.16. We might summarize this by a prior mean of 0.08 and standard deviation of 0.04. These moments correspond to beta hyperparameters of $\alpha_1 = 3.6$ and $\alpha_2 = 41.4$, for which 0.16 is the 96th percentile. Then, the posterior is the $\text{beta}(3.6, 66.4)$, which has mean = 0.051 and 95% posterior equal-tail interval of $(0.013, 0.114)$ and HPD interval of $(0.008, 0.103)$. Stronger prior beliefs result in greater shrinkage of the Bayes estimate toward the prior mean and a narrower posterior equal-tail interval.

1.6.5 Binomial and Multinomial Estimation: Improper Priors

For multinomial data, the sample proportion p_i , is the ML estimate of π_i . It results as the special case of the Bayesian estimate (1.19) when each $\alpha_i = 0$. But when any $\alpha_i = 0$, the Dirichlet formula is not a legitimate probability density function, as it integrates to ∞ instead of 1. It is then an example of an *improper prior distribution*. Bayesian inference sometimes uses such improper prior distributions, as long as the posterior distribution is proper (e.g., Lindley 1964). The Dirichlet posterior is proper as long as $n_i > 0$ for each i having $\alpha_i = 0$.

For parameters that can take value over the entire real line, a common improper distribution is uniform over all real numbers. For a binomial parameter π , the improper beta(0,0) prior for π corresponds to an improper uniform distribution for $\text{logit}(\pi)$. Haldane (1948) suggested that this prior is often sensible in genetics applications, such as for mutation rates for which $\log(\pi)$ might be approximately uniform for π close to 0.

NOTES

Section 1.1: Categorical Response Data

1.1 Measurement scales: Stevens (1951) defined (nominal, ordinal, interval) scales of measurement. Other scales result from mixtures of these types. For instance, *partially ordered* scales occur when subjects respond to questions having categories that are ordered except for don't know or undecided categories.

Section 1.3: Statistical Inference for Categorical Data

1.2 Chi-squared: Greenwood and Nikulin (1996), Kendall and Stuart (1979), and Lancaster (1969) presented in-depth overviews of the chi-squared distribution. Cochran (1952) presented a historical survey of chi-squared tests of fit. See also Cressie and Read (1989), Koch and Bhapkar (1982), Koehler (2005), Moore (1986b), Read and Cressie (1988), and Watson (1959).

1.3 Wald/LR/score: Disadvantages of the Wald method compared with the score and likelihood-ratio methods is that it does not apply when $\hat{\beta}$ is on the boundary of the parameter space (such as a sample proportion $\hat{\pi} = 0$) and its results depend on the parameterization; inference based on $\hat{\beta}$ and its SE is not equivalent to inference based on a nonlinear function of it, such as $\log(\hat{\beta})$ and its SE . See Section 5.2.6. “Higher-order asymptotics” improve on simple normal and chi-squared approximations for distributions of these statistics (Brazzale et al. 2007, Davison et al. 2006).

Section 1.4: Statistical Inference for Binomial Parameters

1.4 Score CI: The superiority of the score interval to the Wald interval for π was shown by, among others, Agresti and Coull (1998), Blyth and Still (1983), Brown et al. (2001), Ghosh (1979), Newcombe (1998a), and Schader and Schmid (1990).

1.5 Continuity correction: Using continuity corrections with large-sample methods provides approximations to exact small-sample methods. We do not present them, since if you prefer an exact method, with modern computational power you can usually implement it directly rather than approximate it. However, we'll see in Sections 3.5.5, 3.5.7, 7.3.7, 16.6.1, and 16.6.4 that exact methods have the disadvantage that they behave conservatively.

1.6 Discreteness: Suppose a statistic T has discrete distribution with cdf $F(t)$. Then, $F(T)$ is *stochastically larger* than uniform over $[0, 1]$, its cdf being everywhere no greater than that of the uniform (Casella and Berger 2001, pp. 77, 434). Likewise, a P -value based on T has null distribution stochastically larger than uniform. In theory, we can eliminate issues with discreteness in tests by performing a supplementary randomization on the boundary of a critical region (see Exercise 1.12). In rejecting H_0 at the boundary with a certain probability, we can obtain type I error probability = α even when α is not an achievable P -value. For such randomization, the P -value is

$$\text{randomized } P\text{-value} = U \times P(T = t_0) + P(T > t_0),$$

where U denotes a uniform $(0, 1)$ random variable (Stevens 1950). In practice, this is not done, as it is absurd to let a random number determine a decision. The mid P -value replaces the arbitrary uniform multiple $U \times P(T = t_0)$ by its expected value $0.50 \times P(T = t_0)$.

Section 1.5: Statistical Inference for Multinomial Parameters

1.7 Multinomials: Other references on testing a specified multinomial include Good et al. (1970) and Baglivo et al. (1992). For simultaneous confidence intervals for multinomial parameters and their differences, see Exercise 1.36, Chafaï (2009), Fitzpatrick and Scott (1987), Goodman (1965), and Sison and Glaz (1995).

Section 1.6: Bayesian Inference for Binomial and Multinomial Parameters

1.8 Beta/Dirichlet priors: Agresti and Hitchcock (2005) surveyed Bayesian methods for categorical data. Lindley (1964) and Good (1965) were influential early articles about Bayesian estimation of multinomial parameters using a Dirichlet prior. Brown et al. (2001) showed that the Jeffreys beta prior yields posterior intervals for the binomial parameter that perform well, having actual coverage probability close to the nominal level. Good (1967) gave a Bayesian goodness-of-fit test that multinomial probabilities are identical, using a hierarchical approach with a symmetric Dirichlet prior that has a log Cauchy distribution for its hyperparameter.

1.9 Loss functions: In decision-theoretic terms, the Bayes estimator minimizes the posterior expected value of a loss function that measures the distance between an estimator $T(\mathbf{y})$ and a parameter θ . It is the posterior mean for squared error loss and posterior median for absolute error loss. For loss function $w(\theta)(T - \theta)^2$, it is $E[\theta w(\theta)|\mathbf{y}]/E[w(\theta)|\mathbf{y}]$. With loss function $(T - \pi)^2/[\pi(1 - \pi)]$ and uniform prior, the Bayes estimator of π is the ML estimator $p = y/n$. Its risk function (the expected loss, treated as a function of π) is constant. Bayes estimators with constant risk are *minimax*, the maximum risk being no greater than the maximum risk for any other estimator. Johnson (1971) showed that p is an admissible estimator, for standard loss functions. For other cases, see DasGupta and Zhang (2004). Blyth (1980) noted that for large n , $E|\hat{\pi} - \pi| \approx \sqrt{2\pi(1 - \pi)/\pi_c n}$, where $\pi_c = 3.14 \dots$ is the mathematical constant.

EXERCISES

Applications

1.1 Identify each variable as nominal, ordinal, or interval.

- a. UK political party preference (Labour, Liberal Democrat, Conservative)
- b. Anxiety rating (none, mild, moderate, severe, very severe)
- c. Patient survival (in number of months)
- d. Clinic location (London, Boston, Madison, Rochester, Montreal)
- e. Response of tumor to chemotherapy (complete elimination, partial reduction, stable, growth progression)
- f. Favorite grocery store for UK residents (Sainsbury, Tesco, Waitrose, other)

1.2 Each of 100 multiple-choice questions on an exam has four possible answers, one of which is correct. For each question, a student guesses by selecting an answer randomly.

- a. Specify the distribution of the number of correct answers.
- b. Find the mean and standard deviation of that distribution. Would it be surprising if the student made at least 50 correct responses? Why?
- c. Specify the distribution of (n_1, n_2, n_3, n_4) , where n_j is the number of times the student picked choice j .
- d. Find $E(n_j)$ and $\text{var}(n_j)$. Show that $\text{cov}(n_j, n_k) = -6.25$ and $\text{corr}(n_j, n_k) = -0.333$.

1.3 An experiment studies the number of insects that survive a certain dose of an insecticide, using several batches of insects of size n each. The insects are sensitive to factors that vary among batches during the experiment but were not measured, such as temperature level. Explain why the distribution of the number of insects per batch surviving the experiment might show overdispersion relative to a $\text{bin}(n, \pi)$ distribution.

1.4 In his autobiography *A Sort of Life*, British author Graham Greene described a period of severe mental depression during which he played Russian roulette. This “game” consists of putting a bullet in one of the six chambers of a pistol, spinning the chambers to select one at random, and then firing the pistol once at one’s head.

- a. Greene played this game six times and was lucky that none of them resulted in a bullet firing. Find the probability of this outcome.
- b. Suppose that he had kept playing this game until the bullet fired. Let Y denote the number of the game on which it fires. Explain why the probability mass function for Y is the *geometric*, $p(y) = (5/6)^{y-1} (1/6)$, $y = 1, 2, 3, \dots$.

1.5 When the 2010 General Social Survey asked, “Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if she is married and does not want any more children,” 587 replied “yes” and 636 replied “no.” Let π denote the population proportion who would reply “yes.” Find the P -value for testing $H_0: \pi = 0.50$ using the score test, and construct a 95% confidence interval for π . Interpret the results.

1.6 Refer to the vegetarianism example in Section 1.4.3. For testing $H_0: \pi = 0.50$ against $H_a: \pi \neq 0.50$, show that:

- a. The likelihood-ratio statistic equals $2[25 \log(25/12.5)] = 34.7$.
- b. The chi-squared form of the score statistic equals 25.0.
- c. The Wald z or chi-squared statistic is infinite.

1.7 In a crossover trial comparing a new drug to a standard, π denotes the probability that the new one is judged better. It is desired to estimate π and test $H_0: \pi = 0.50$ against $H_a: \pi \neq 0.50$. In 20 independent observations, the new drug is better each time.

- a. Find and sketch the likelihood function. Is it close to the quadratic shape that large-sample

normal approximations utilize?

- b.** Give the ML estimate of π . Conduct a Wald test and construct a 95% Wald confidence interval for π . Are these sensible?
- c.** Conduct a score test, reporting the P -value. Construct a 95% score confidence interval. Interpret.
- d.** Conduct a likelihood-ratio test and construct a likelihood-based 95% confidence interval. Interpret.
- e.** Construct an exact binomial test. Interpret.

1.8 Refer to the previous exercise. Suppose you wanted a large enough sample to estimate the probability of preferring the new drug to within 0.05, with confidence 0.95. If the true probability is 0.80, about how large a sample is needed?

1.9 In an experiment on chlorophyll inheritance in maize, for 1103 seedlings of self-fertilized heterozygous green plants, 854 seedlings were green and 249 were yellow. Theory predicts the ratio of green to yellow is 3:1. Test the hypothesis that 3:1 is the true ratio. Report the P -value, and interpret.

1.10 [Table 1.3](#) contains Ladislaus von Bortkiewicz's data on deaths of soldiers in the Prussian army from kicks by army mules (Fisher 1934, Quine and Seneta 1987). The data refer to 10 army corps, each observed for 20 years. In 109 corps-years of exposure, there were no deaths, in 65 corps-years there was one death, and so on. Estimate the mean and test whether probabilities of occurrences in these five categories follow a Poisson distribution (truncated for 4 and above).

Table 1.3 Data on Deaths by Mule Kicks, for Exercise 1.10

Number of Deaths	Number of Corps-Years
0	109
1	65
2	22
3	3
4	1
≥ 5	0

1.11 A binomial experiment tests $H_0: \pi = 0.50$ against $H_a: \pi \neq 0.50$ using significance level 0.05. Only $n = 5$ observations are available. Show that the true null probability of rejecting H_0 is 0.00 for an exact binomial test and $\frac{1}{16}$ using the large-sample score test.

1.12 A researcher routinely tests using a nominal $P(\text{type I error}) = 0.05$, rejecting H_0 if the P -value ≤ 0.05 . An exact test using test statistic T has null distribution $P(T = 0) = 0.30$, $P(T = 1) = 0.62$, and $P(T = 2) = 0.08$, where a higher T provides more evidence against the null.

- a.** With the usual P -value, show that the actual $P(\text{type I error}) = 0$.
- b.** With the mid P -value, show that the actual $P(\text{type I error}) = 0.08$.
- c.** Find $P(\text{type I error})$ in parts (a) and (b) when $P(T = 0) = 0.30$, $P(T = 1) = 0.66$, $P(T = 2) = 0.04$. Note that the test with mid P -value can be conservative or liberal. The exact test with ordinary P -value cannot be liberal.
- d.** In part (a), a randomized-decision test generates a uniform random variable U from $[0, 1]$ and rejects H_0 if both $T = 2$ and $U \leq \frac{5}{8}$. Show the actual $P(\text{type I error}) = 0.05$. Is this a sensible test?

1.13 The 2006 General Social Survey asked respondents how much government should spend on culture and the arts, with categories (much more, more, the same, less, much less). For 18–21 year-old females, the counts in these categories were (0, 8, 10, 9, 1). Find the Bayes estimates of the population proportions based on a Dirichlet prior distribution with $\{\alpha_i = K / 5\}$ for values of $K = 1, 2.5, 5$. For each case, compare the estimate for the “much more” category to the ML estimate.

1.14 Refer to Example 1.6.4 on estimating the proportion of vegetarians. For the Jeffreys prior,

find the posterior mean, the posterior 95% equal-tail interval, and the 95% highest posterior density interval.

1.15 You plan to use Bayesian methods to estimate binomial parameters in two cases, using n observations. In case (1) you want to estimate the probability that a new treatment for skin cancer is effective. In case (2) you want to estimate the probability of a head when you repeatedly flip a particular coin. Select prior distributions that you think would be sensible for each case. If they differ, explain why.

Theory and Methods

1.16 It is easier to get a precise estimate of the binomial parameter when π is near 0 or 1 than when it is near $\frac{1}{2}$. Explain why.

1.17 Suppose that $P(Y_i = 1) = 1 - P(Y_i = 0) = \pi$, $i = 1, \dots, n$, where $\{Y_i\}$ are independent. Let $Y = \sum_i Y_i$.

- a. What is the distribution of Y ? What are $E(Y)$ and $\text{var}(Y)$?
- b. When $\{Y_i\}$ instead have pairwise correlation $\rho > 0$, show that $\text{var}(Y) > n\pi(1 - \pi)$, overdispersion relative to the binomial. [Altham (1978) and Ochi and Prentice (1984) discussed generalizations of the binomial that allow correlated trials.]
- c. Suppose that heterogeneity exists: $P(Y_i = 1|\pi) = \pi$ for all i , but π is a random variable with density function $g(\cdot)$ on $[0, 1]$ having mean ρ and positive variance. Show that $\text{var}(Y) > n\rho(1 - \rho)$. (When π has a beta distribution, Y has the *beta-binomial distribution* of Section 14.3.)

1.18 For a sequence of independent Bernoulli trials, let Y be the number of successes before the k th failure. Explain why its probability mass function is the *negative binomial*,

$$p(y) = \frac{(y+k-1)!}{y!(k-1)!} \pi^y (1-\pi)^{k-y}, \quad y = 0, 1, 2, \dots$$

[For it, $E(Y) = k\pi/(1 - \pi)$ and $\text{var}(Y) = k\pi/(1 - \pi)^2$, so $\text{var}(Y) > E(Y)$; the Poisson is the limit as $k \rightarrow \infty$ and $\pi \rightarrow 0$ with $k\pi = \mu$ fixed.]

1.19 For the multinomial distribution, show that

$$\text{corr}(n_j, n_k) = -\pi_j \pi_k / \sqrt{\pi_j(1 - \pi_j)\pi_k(1 - \pi_k)}.$$

When $c = 2$, show that this simplifies to $\text{corr}(n_1, n_2) = -1$, and explain why this makes intuitive sense.

1.20 Show that the moment generating function (mgf) is (a) $m(t) = (1 - \pi + \pi e_t)^n$ for the binomial distribution, (b) $m(t) = \exp\{\mu[\exp(t) - 1]\}$ for the Poisson distribution. For each distribution, use them to obtain the first two moments and to show a reproductive property.

1.21 A likelihood-ratio statistic equals t_o . At the ML estimates, show that the data are $\exp(t_o/2)$ times more likely under H_a than under H_0 .

1.22 Suppose that y_1, y_2, \dots, y_n are independent from a Poisson distribution.

- a. Obtain the likelihood function. Show that the ML estimator $\hat{\mu}_o = \bar{y}$.
- b. Construct a large-sample test statistic for $H_0: \mu = \mu_0$ using (i) the Wald method, (ii) the score method, and (iii) the likelihood-ratio method.
- c. Explain how to construct a large-sample confidence interval for μ using (i) the Wald method, (ii) the score method, and (iii) the likelihood-ratio method.

1.23 Inference for Poisson parameters can often be based on connections with binomial and multinomial distributions. Show how to test $H_0: \mu_1 = \mu_2$ for two populations based on independent Poisson counts (y_1, y_2) , using a corresponding binomial test. [Hint: Condition on $n = y_1 + y_2$ and identify $\pi = \mu_1/(\mu_1 + \mu_2)$.] How can you construct a confidence interval for μ_1/μ_2 based on one for π ?

1.24 Since the Wald confidence interval for a binomial parameter π is degenerate when $\pi = 0$ or 1, argue that the probability that the interval covers π cannot exceed $[1 - \pi^n - (1 - \pi)^n]$; hence,

the infimum of the coverage probability over $0 < \pi < 1$ equals 0, regardless of n .

1.25 We noted in Section 1.4.2 that the midpoint $\tilde{\pi}$ of the score confidence interval (1.14) for π is the sample proportion after adding $z_{\alpha/2}^2$ observations to the sample, half of each type. This motivates a simple confidence interval,

$$\tilde{\pi} \pm z_{\alpha/2} \sqrt{\tilde{\pi}(1 - \tilde{\pi})/n^*}, \quad \text{where } n^* = n + z_{\alpha/2}^2.$$

Show that the variance $\tilde{\pi}(1 - \tilde{\pi})/n^*$ at the weighted average is at least as large as the weighted average of the variances that appears under the square root sign in the score interval. [Hint: Use Jensen's inequality.] Thus, this interval, which is sometimes referred to as the *Agresti–Coull confidence interval*, contains the score interval. [Agresti and Coull (1998) and Brown et al. (2001) showed that it performs much better than the Wald interval. It does not have the score interval's disadvantage (Exercise 16.32) of poor coverage near 0 and 1. With 95% confidence, this motivates a simple method that uses the Wald method after adding 2 observations of each type (Agresti and Coull 1998, Agresti and Caffo 2000); this is sometimes called the *plus four confidence interval*.]

1.26 A binomial sample of size n has $y = 0$ successes.

- a. Show that the confidence interval for π based on the likelihood function is $[0.0, 1 - \exp(-z_{\alpha/2}^2/2n)]$. For $\alpha = 0.05$, use the expansion of an exponential function to show that this is approximately $[0, 1.92/n]$.
- b. For the score method, show that the confidence interval is $[0, z_{\alpha/2}^2/(n + z_{\alpha/2}^2)]$, or $[0, 3.84/(n + 3.84)]$ when $\alpha = 0.05$. (See Exercise 16.30 for small-sample intervals when $y = 0$.)

1.27 Suppose that $P(T = t_j) = \pi_j, j = 1, \dots$. Show that $E(\text{mid } P\text{-value}) = 0.50$. [Hint: Show that $\sum_j \pi_j(\pi_j/2 + \pi_{j+1} + \dots) = (\sum_j \pi_j)^2/2$.]

1.28 For a statistic T with cdf $F(t)$ and $p(t) = P(T = t)$, the *mid distribution function* is $F_{\text{mid}}(t) = F(t) - 0.50p(t)$ (Parzen 1997). Given $T = t_o$, show that the mid P -value equals $1 - F(t_o)$. (It also satisfies $E[F_{\text{mid}}(T)] = 0.50$ and $\text{var}[F_{\text{mid}}(T)] = (1/12)\{1 - E[p^2(T)]\}$.)

1.29 Genotypes AA, Aa, and aa occur with probabilities $[\theta^2, 2\theta(1 - \theta), (1 - \theta)^2]$. A multinomial sample of size n has frequencies (n_1, n_2, n_3) of these three genotypes.

- a. Form the log likelihood. Show that $\theta = (2n_1 + n_2)/(2n_1 + 2n_2 + 2n_3)$.
- b. Show that $-\partial^2 L(\theta)/\partial\theta^2 = [(2n_1 + n_2)/\theta^2] + [(n_2 + 2n_3)/(1 - \theta)^2]$ and that its expectation is $2n/\theta(1 - \theta)$. Use this to obtain an asymptotic standard error of θ .
- c. Explain how to test whether the probabilities truly have this pattern.

1.30 Refer to Section 1.5.6 and the model for pneumonia infections in calves. Using the likelihood function to obtain the information, show that the approximate standard error of $\tilde{\pi}$ is $\sqrt{\pi(1 - \pi)/n(1 + \pi)}$.

1.31 Refer to Section 1.5.6. Let a denote the number of calves that got a primary, secondary, and tertiary infection, b the number that received a primary and secondary but not a tertiary infection, c the number that received a primary but not a secondary infection, and d the number that did not receive a primary infection. Let π be the probability of a primary infection. Consider the hypothesis that the probability of infection at time t , given infection at times $1, \dots, t - 1$, is also π , for $t = 2, 3$. Show that $\hat{\pi} = (3a + 2b + c)/(3a + 3b + 2c + d)$.

1.32 Refer to quadratic form (1.18) that leads to the Pearson chi-squared.

- a. Verify that the matrix quoted in the text for Σ_0^{-1} is the inverse of Σ_0 .
- b. Show that (1.18) simplifies to Pearson's statistic (1.16).
- c. For the z_s statistic (1.11), show that $z_s^2 = X^2$ for $c = 2$.

1.33 For testing $H_0: \pi_j = \pi_{j0}, j = 1, \dots, c$, using sample multinomial proportions $\{\hat{\pi}_j\}$, the likelihood-ratio statistic (1.17) is

$$G^2 = -2n \sum_j \hat{\pi}_j \log(\pi_{j0}/\hat{\pi}_j).$$

Show that $G^2 \geq 0$, with equality if and only if $\hat{\pi}_j = \pi_{j0}$ for all j . [Hint: Apply Jensen's inequality to $E(-2n \log X)$, where X equals $\pi_{j0}/\hat{\pi}_j$ with probability $\hat{\pi}_j$.]

1.34 For counts $\{n_i\}$, the *power divergence statistic* for testing goodness of fit (Cressie and Read 1984, Read and Cressie 1988) is

$$\frac{2}{\lambda(\lambda+1)} \sum n_i [(n_i/\hat{\mu}_i)^\lambda - 1] \quad \text{for } -\infty < \lambda < \infty.$$

- a. For $\lambda = 1$, show that this equals χ^2 .
- b. As $\lambda \rightarrow 0$, show that it converges to G^2 . [Hint: $\log t = \lim_{h \rightarrow 0} (t^h - 1)/h$.]
- c. As $\lambda \rightarrow -1$, show that it converges to $2 \sum \hat{\mu}_i \log(\hat{\mu}_i / n_i)$, the *minimum discrimination information* statistic (Gokhale and Kullback 1978).
- d. For $\lambda = -2$, show that it equals $\Lambda(n_i - \hat{\mu}_i)^2 / n_i$, the *Neyman modified chi-squared* statistic (Neyman 1949).
- e. For $\lambda = -\frac{1}{2}$, show that it equals $4 \sum (\sqrt{n_i} - \sqrt{\hat{\mu}_i})^2$ the *Freeman–Tukey* statistic (Freeman and Tukey 1950).

[Under regularity conditions, their asymptotic distributions are identical (Drost et al. 1989). The chi-squared null approximation works best for Λ near $\frac{2}{3}$.]

1.35 The chi-squared mgf with $\text{df} = v$ is $m(t) = (1 - 2t)^{-v/2}$, for $|t| < \frac{1}{2}$. Use it to prove the reproductive property of the chi-squared distribution.

1.36 For the multinomial $(n, \{\pi_j\})$ distribution with $c > 2$, a possible set of score-type simultaneous confidence limits for π_j , are the solutions of

$$(\hat{\pi}_j - \pi_j)^2 / [\pi_j(1 - \pi_j)/n] = (z_{\alpha/2c})^2, \quad j = 1, \dots, c.$$

- a. Using the Bonferroni inequality, argue that for large n these c intervals simultaneously contain all $\{\pi_j\}$ with probability at least $1 - \alpha$.
- b. Show that the standard deviation of $\hat{\pi}_j - \hat{\pi}_k$ is $[\pi_j + \pi_k - (\pi_j - \pi_k)^2]/n$. Let $a = c(c-1)/2$. For large n , explain why the probability is at least $1 - \alpha$ that the Wald confidence intervals $(\hat{\pi}_j - \hat{\pi}_k) \pm z_{\alpha/2a} \{[\hat{\pi}_j + \hat{\pi}_k - (\hat{\pi}_j - \hat{\pi}_k)^2]/n\}^{1/2}$ simultaneously contain the a differences $\{\pi_j - \pi_k\}$ (Goodman 1965).

1.37 Consider the Bayesian equal-tail posterior interval for a binomial parameter π , using a beta or logit-normal prior. When $y = 0$, explain why the lower limit for π can never be 0, unlike the frequentist approach based on inverting a score or likelihood-ratio test.

1.38 Consider estimating the ratio π_i/π_j of two multinomial parameters. Should the estimate depend at all on the counts in other categories?

- a. With a frequentist approach, explain why the ML estimate of π_i/π_j is n_i/n_j .
- b. For a Dirichlet prior, show that using the Bayes estimates of π_i and π_j to estimate π_i/π_j uses also the counts in other categories. (However, the posterior distribution of $\gamma = \pi_i/(\pi_i + \pi_j)$ is the same as its posterior distribution ignoring the other counts and treating y_i as binomial with sample size $(y_i + y_j)$ and parameter γ .)

1.39 Given π , Y has a $\text{bin}(n, \pi)$ distribution, and π has a uniform prior distribution. Show that the marginal distribution of Y is uniform over $0, 1, \dots, n$.

1.40 Consider the Bayes estimator of the binomial parameter π using a beta prior distribution.

- a. Show that the ML estimator is a limit of Bayes estimators, for a certain sequence of beta prior parameter values.
- b. Find an improper prior density such that the Bayes estimator coincides with the ML estimator. (In this sense, the ML estimator is a *generalized Bayes estimator*.)

1.41 For the Dirichlet prior for multinomial probabilities, show the posterior expected value of

π_i is formula (1.19). Derive the expression for this Bayes estimator as a weighted average of p_i and $E(\pi_i)$.

¹See www.stat.tamu.edu/~west/applets/binomialdemo2.html.

²The T superscript on a vector or matrix denotes the transpose.

³See logitnorm.r-forge.r-project.org and the “Logit-normal distribution” entry in wikipedia.org for figures illustrating the shapes described below.

CHAPTER 2

Describing Contingency Tables

In this chapter we introduce parameters that summarize tables displaying relationships between categorical variables. After introducing basic terminology and notation in Section 2.1, in Section 2.2 we introduce measures for comparing two groups on a categorical response. The *odds ratio* has special importance, appearing as a parameter in models discussed later. In Section 2.3 we extend the scope by controlling for a third variable. The association can change dramatically under a control. The chapter's primary focus is binary variables, but in Section 2.4 we present parameters for nominal and ordinal variables.

2.1 PROBABILITY STRUCTURE FOR CONTINGENCY TABLES

Let X and Y denote two categorical variables, X with I categories and Y with J categories. Classifications of subjects on both variables have IJ possible combinations. When both variables are response variables, we focus on their *joint distribution*, which also determines the *marginal and conditional distributions*. When Y is a response variable and X is an explanatory variable, we focus on the *conditional distribution* of Y and how it changes as the category of X changes.

2.1.1 Contingency Tables

A rectangular table having I rows for categories of X and J columns for categories of Y displays the IJ possible combinations of outcomes. The *cells* of the table represent the IJ possible outcomes. When the cells contain frequency counts of outcomes for a sample, the table is called a *contingency table*, a term introduced by Karl Pearson (1904). Another name is *cross-classification table*. A contingency table with I rows and J columns is called an I -by- J (denoted by $I \times J$) table.

[Table 2.1](#), a 2×3 contingency table, is from a report on the relationship between aspirin use and heart attacks by the Physicians' Health Study Research Group at Harvard Medical School. The Physicians' Health Study was a 5-year randomized study of whether regular aspirin intake reduces mortality from cardiovascular disease. Every other day, physicians participating in the study took either one aspirin tablet or a placebo. The study was *blind*—those in the study did not know whether they were taking aspirin or a placebo. Of the 11,034 physicians taking a placebo, 18 suffered fatal heart attacks over the course of the study, whereas of the 11,037 taking aspirin, 5 had fatal heart attacks.

Table 2.1 Cross-Classification of Aspirin Use and Myocardial Infarction

Myocardial Infarction			
	Fatal Attack	Nonfatal Attack	No Attack
Placebo	18	171	10,845
Aspirin	5	99	10,933

Source: Preliminary report: Findings from the aspirin component of the ongoing Physicians' Health Study. *N. Engl. J. Med.* **318**: 262–264, 1988.

2.1.2 Joint/Marginal/Conditional Distributions for Contingency Tables

In some applications, both X and Y are response variables. Suppose subjects are randomly chosen from a particular population, such as in a sample survey employing simple random sampling. Then, the responses (X, Y) of a randomly chosen subject have a probability distribution. Let π_{ij} denote the probability that (X, Y) occurs in the cell in row i and column j . The probability distribution $\{\pi_{ij}\}$ is the *joint distribution* of X and Y . The *marginal distributions* are the row and column totals that result from summing the joint probabilities. We denote these by $\{\pi_{i+}\}$ for the row variable and $\{\pi_{+j}\}$ for the column variable, where the subscript “ $+$ ” denotes the sum over that index; that is,

$$\pi_{i+} = \sum_j \pi_{ij} \quad \text{and} \quad \pi_{+j} = \sum_i \pi_{ij}.$$

These satisfy $\sum_i \pi_{i+} = \sum_j \pi_{+j} = \sum_i \sum_j \pi_{ij} = 1.0$. The marginal distributions provide single-variable information.

In most contingency tables, [Table 2.1](#) being an example, one variable—say, Y —is a response variable and the other (X) is an explanatory variable. When X is fixed rather than random, the notion of a joint distribution for X and Y is no longer meaningful. However, for a fixed category of X , Y has a probability distribution. It is germane to study how this distribution changes as the category of X changes. Given that a subject is classified in row i of X , we use $\pi_{j|i}$ to denote the probability of classification in column j of Y , $j = 1, \dots, J$. Then, $\sum_j \pi_{j|i} = 1$. The probabilities $\{\pi_{1|i}, \dots, \pi_{J|i}\}$ form the *conditional distribution* of Y at category i of X . A principal aim of many studies is to compare conditional distributions of Y at various levels of explanatory variables.

When both variables are response variables, descriptions of the association can use their joint distribution, the conditional distribution of Y given X , or the conditional distribution of X given Y . The conditional distribution of Y given X relates to the joint distribution by

$$\pi_{j|i} = \pi_{ij}/\pi_{i+} \quad \text{for all } i \text{ and } j.$$

[Table 2.2](#) displays notation for joint, conditional, and marginal distributions for the 2×2 case. Sample distributions use similar notation, with p or $\hat{\pi}$ in place of π . For instance, $\{p_{ij}\}$ denotes the sample joint distribution. The cell frequencies are denoted by $\{n_{ij}\}$, and $n = \sum_i \sum_j n_{ij}$ is the total sample size. Thus,

Table 2.2 Notation for Joint, Conditional, and Marginal Probabilities

Row	Column		Total
	1	2	
1	π_{11} ($\pi_{1 1}$)	π_{12} ($\pi_{2 1}$)	π_{1+} (1.0)
2	π_{21} ($\pi_{1 2}$)	π_{22} ($\pi_{2 2}$)	π_{2+} (1.0)
Total	π_{+1}	π_{+2}	1.0

$$p_{ij} = n_{ij}/n.$$

The sample proportion of times that subjects in row i made response j is $p_{j|i} = p_{ij}/P_{i+} = n_{ij}/n_{i+}$, where $n_{i+} = np_{i+} = \sum_j n_{ij}$.

2.1.3 Example: Sensitivity and Specificity for Medical Diagnoses

Diagnostic tests are used to help detect certain medical conditions. These include the PSA blood test for prostate cancer and imaging devices such as the mammogram for diagnosing breast cancer and X-rays and the MRI body scan. A diagnostic test for a condition is said to be *positive* if it states that the condition is present and *negative* if it states that the condition is absent.

Breast cancer is the most common form of cancer in women, affecting about 10% at some time in their lives. For the mammogram diagnostic test, the chance of a correct test result varies according to the breast density and the radiologist's level of experience. Let X = true disease status (i.e., whether a woman truly has breast cancer) and let Y = diagnosis (positive, negative). [Table 2.3](#) shows typically reported values for conditional probabilities of Y given X .

Table 2.3 Estimated Conditional Distributions for Breast Cancer Mammograms

Breast Cancer	Diagnosis of Test		
	Positive	Negative	Total
Yes	0.86	0.14	1.0
No	0.12	0.88	1.0

With a diagnostic test, the two correct diagnoses are a positive outcome when the person has the disease and a negative outcome when a person does not have it. Given that the person has the disease, the conditional probability that the test is positive is called the *sensitivity*. Given that the person does not have the disease, the conditional probability that the test is negative is called the *specificity* (Yerushalmy 1947). Ideally, these are both very high.

For a 2×2 table with the format of [Table 2.3](#), sensitivity is $\pi_{1|1}$ and specificity is $\pi_{2|2}$. In [Table 2.3](#), the estimated sensitivity of mammography is 0.86. Of women with breast cancer, 86% are diagnosed correctly. The estimated specificity is 0.88. Of women not having breast cancer, 88% are diagnosed correctly.

2.1.4 Independence of Categorical Variables

Two categorical response variables are defined to be *independent* if all joint probabilities equal the product of their marginal probabilities,

$$(2.1) \quad \pi_{ij} = \pi_{i+}\pi_{+j} \quad \text{for } i = 1, \dots, I \quad \text{and} \quad j = 1, \dots, J.$$

When X and Y are independent,

$$\pi_{j|i} = \pi_{ij}/\pi_{i+} = (\pi_{i+}\pi_{+j})/\pi_{i+} = \pi_{+j} \quad \text{for } i = 1, \dots, I.$$

Each conditional distribution of Y is identical to the marginal distribution of Y .

Thus, two variables are independent when $\{\pi_{j|1} = \dots = \pi_{j|I}, \text{ for } j = 1, \dots, J\}$; that is, the probability of any given column response is the same in each row. When Y is a response and X is an explanatory variable, this is a more natural way to define independence than (2.1). Independence is then often referred to as *homogeneity* of the conditional distributions.

2.1.5 Poisson, Binomial, and Multinomial Sampling

The probability distributions introduced in Section 1.2 extend to cell counts in contingency tables. For instance, a Poisson sampling model treats cell counts $\{Y_{ij}\}$ as independent Poisson random variables with parameters $\{\mu_{ij}\}$. The joint probability mass function for potential outcomes $\{n_{ij}\}$ is then the product of the Poisson probabilities $P(Y_{ij} = n_{ij})$ for the IJ cells, or

$$\text{Poisson sampling: } \prod_i \prod_j \exp(-\mu_{ij}) \mu_{ij}^{n_{ij}} / n_{ij}!.$$

When the total sample size n is fixed but the row and column totals are not, a *multinomial sampling* model applies. The IJ cells are the possible outcomes. The probability mass function of the cell counts has the multinomial form

$$\text{multinomial sampling: } [n!/(n_{11}! \cdots n_{IJ}!)] \prod_i \prod_j \pi_{ij}^{n_{ij}}.$$

When observations on a response Y occur separately at each setting of an explanatory variable X , it is natural to treat row totals as fixed. For simplicity, we then use the notation $n_i = n_{i+}$. Suppose that the n_i observations on Y at setting i of X are independent, each with probability distribution $\{\pi_{1|i}, \dots, \pi_{J|i}\}$. The counts $\{n_{ij}, j = 1, \dots, J\}$ satisfying $\sum_j n_{ij} = n_i$ then have multinomial form. When samples at different settings of X are independent, the joint probability function for the entire data set is the product of the multinomial functions from the various settings. This sampling scheme is *independent multinomial sampling*,

$$(2.2) \quad \text{independent multinomial sampling: } \prod_i \left[\frac{n_i!}{\prod_j n_{ij}!} \prod_j \pi_{j|i}^{n_{ij}} \right],$$

also called *product multinomial sampling*. The special case $J = 2$ is *independent binomial sampling*.

Independent multinomial sampling also results under the following conditions: Suppose that $\{n_{ij}\}$ result from either independent Poisson sampling with means $\{\mu_{ij}\}$ or multinomial sampling over the IJ cells with probabilities $\{\pi_{ij} = \mu_{ij}/n\}$. When X is an explanatory variable, it is sensible to perform statistical inference conditional on the totals $\{n_{i+} = \sum_j n_{ij}\}$ even when their values are not fixed by the sampling design. Conditional on $\{n_i\}$, the cell counts $\{n_{ij}, j = 1, \dots, J\}$ have the multinomial distribution (2.2) with response probabilities $\{\pi_{j|i} = \mu_{ij} / \mu_{i+}, j = 1, \dots, J\}$, and cell counts from different rows are independent. With this conditioning, we treat the row totals as fixed and analyze the data as if they formed separate independent samples.

Sometimes both row and column margins are naturally fixed. The appropriate sampling distribution is then usually the *hypergeometric*. This case, considered in Section 3.5.1, is less common.

2.1.6 Example: Seat Belts and Auto Accident Injuries

Researchers in the Massachusetts Department of Transportation (MassDOT) plan to study the effects of cell-phone use and seat-belt use on incidence and severity of traffic accidents. For the relationship between seat-belt use (yes, no) and outcome of an automobile accident (fatality, nonfatality) for drivers involved in accidents on the Massachusetts Turnpike, they will summarize results in the format shown in [Table 2.4](#). They plan to catalog all accidents on the turnpike for the next year, classifying each according to these variables. The total sample size is then a random variable. They might treat the numbers of observations at the four combinations of seat-belt use and outcome of crash as independent Poisson random variables with unknown means $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}\}$.

Table 2.4 Seat-Belt Use and Results of Automobile Accidents

		Result of Accident	
		Fatality	Nonfatality
Seat-Belt Use	Yes		
	No		

Suppose, instead, that the researchers randomly sample 200 police records of accidents on the turnpike in the past year and classify each according to seat-belt use and outcome of the accident. For this study, the total sample size n is fixed. They might then treat the four cell counts as a multinomial random variable with $n = 200$ trials and unknown joint probabilities $\{\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}\}$.

Suppose, instead, that police records for accidents involving fatalities were filed separately from the others. The researchers might instead randomly sample 100 records of accidents with a fatality and randomly sample 100 records of accidents with no fatality. This approach fixes the column totals in [Table 2.4](#) at 100. They might then regard each column of [Table 2.4](#) as an independent binomial sample. Yet another approach, the traditional experimental design, takes 200 subjects and randomly assigns 100 of them to wear seat belts and the other 100 not to wear them; then the 200 all are forced to have an accident. The recorded results would then be independent binomial samples in each row, with fixed row totals of 100 each. (Obviously, traditional designs common in some experimental science may not be ethical for humans, especially in some medical research.)

2.1.7 Example: Case–Control Study of Cancer and Smoking

[Table 2.5](#) comes from one of the first studies of the link between lung cancer and smoking. Richard Doll and Austin Bradford Hill investigated this with data from 20 hospitals in London, England, at a time when many medical scientists thought that the increasing rates of lung cancer in London mainly reflected increasing air pollution, largely from the burning of coal (and thus, the frequent “London fog”) before the Clean Air Act of 1956. In their study, patients admitted with lung cancer in the preceding year were queried about their smoking behavior. For each of the 709 patients admitted, they recorded the smoking behavior of a noncancer patient at the same hospital of the same gender and within the same 5-year grouping on age. The 709 *cases* in the first column of [Table 2.5](#) are those having lung cancer and the 709 *controls* in the second column are those not having it. A smoker was defined as a person who had smoked at least one cigarette a day for at least a year.

Table 2.5 Cross-Classification of Smoking by Lung Cancer

Smoker	Lung Cancer	
	Cases	Controls
Yes	688	650
No	21	59
Total	709	709

Source: Based on data reported in Table IV, R. Doll and A. B. Hill, *Br. Med. J.*, 739–748, Sept. 30, 1950.

Normally, whether lung cancer occurs is a response variable and smoking behavior is an explanatory variable. In this study, however, the marginal distribution of lung cancer is fixed by the sampling design, and the outcome measured is whether the subject ever was a smoker. The study, which uses a *retrospective* design to “look into the past,” is called a *case-control study*. Such studies are common in health-related applications. Often, the two samples are matched, as in this study. Sometimes the samples of cases and controls are independent rather than matched. For instance, another early case–control study on lung cancer and smoking sampled subjects by sending letters to the estates of physicians who had died of some type of cancer in 1950 or 1951, and observations were cross-classified on type of cancer and the subject’s smoking behavior (Cornfield 1956).

We might want to compare smokers with nonsmokers in terms of the proportion who suffered lung cancer. These proportions refer to the conditional distribution of lung cancer, given smoking behavior. Instead, case-control studies provide proportions in the reverse direction, for the conditional distribution of smoking behavior, given lung cancer status. For those in [Table 2.5](#) with lung cancer, the proportion who were smokers was $688/709 = 0.970$, while it was $650/709 = 0.917$ for the controls.

When we know the proportion of the population having lung cancer, we can use Bayes’ theorem to compute sample conditional distributions in the direction of main interest (Exercise 2.25). Otherwise, using a retrospective sample, we cannot estimate the probability of lung cancer at each category of smoking behavior. For [Table 2.5](#) we do not know the population prevalence of lung cancer, and the patients suffering it were probably sampled at a rate far in excess of their occurrence in the general population.

2.1.8 Types of Studies: Observational Versus Experimental

By contrast to the case-control study just described, imagine a study that samples subjects from the population of teenagers and then 60 years later measures the rates of lung cancer for the smokers and nonsmokers. Such a sampling design is *prospective*. There are two types of prospective studies. *Clinical trials* randomly allocate subjects to the groups who will be smokers and nonsmokers. In *cohort studies*, subjects make their own choice about whether to smoke, and the study observes in future time who develops lung cancer. Yet another approach, a *cross-sectional design*, samples subjects and classifies them simultaneously on both variables.

Prospective studies usually condition on the totals $\{n_i = \sum_j n_{ij}\}$ for categories of X and regard each row of J counts as an independent multinomial sample on Y . *Retrospective studies* treat the totals $\{n_{+j}\}$ for Y as fixed and regard each column of I counts as a multinomial sample on X . In *cross-sectional studies*, the total sample size is fixed but not the row or column totals, and the IJ cell counts are a multinomial sample.

A clinical trial is an *experimental* study, the investigator having the advantage of experimental control over which subjects receive each treatment. Such studies can use the power of randomization to make the groups balance (apart from sampling error) on other variables that may be associated with the response. This lowers the chance that an association may be due to some unobserved variable. By contrast, case-control, cohort, and cross-sectional studies are *observational* studies. They merely observe who chooses each group and who has the outcome of interest. Observational studies have more potential for biases of various types, and it is dangerous to conclude that an association reflects a causal connection.

For example, suppose an observational study finds that people who are unmarried are more likely to be a member of Facebook than those who are married. Many variables are associated both with marital status and with whether a person is a member of Facebook. Such variables could account for the association. One such variable could be a person's age. Perhaps younger people are both more likely to be a member of Facebook and more likely to be unmarried. If the study failed to measure age or control for it adequately, it might misleadingly predict a causal relation between marital status and Facebook membership.

2.2 COMPARING TWO PROPORTIONS

Many studies are designed to compare groups on a binary response variable. Then Y has only two categories, such as (success, failure) for outcome of a medical treatment. With two groups, a 2×2 contingency table displays the results. The rows are the groups and the columns are the categories of Y . This section presents parameters for comparing the groups.

2.2.1 Difference of Proportions

For subjects in row i , $\pi_{1|i}$ is the probability that the response has outcome in category 1 (“success”). With only two possible outcomes, $\pi_{2|i} = 1 - \pi_{1|i}$, and we use the simpler notation π_i for $\pi_{1|i}$. The *difference of proportions* of successes, $\pi_1 - \pi_2$, is a basic comparison of the two rows. Comparison on failures is equivalent to comparison on successes, since

$$(1 - \pi_1) - (1 - \pi_2) = \pi_2 - \pi_1.$$

The difference of proportions falls between -1.0 and $+1.0$. It equals zero when the rows have identical conditional distributions. The response Y is independent of the row classification when $\pi_1 - \pi_2 = 0$.

When both variables are responses, conditional distributions apply in either direction. We can also compare the two columns, such as by the difference between the proportions in row 1. This usually is not equal to the difference $\pi_1 - \pi_2$ comparing the rows, unless $\pi_1 - \pi_2 = 0$.

2.2.2 Relative Risk

A value $\pi_1 - \pi_2$ of fixed size may have greater importance when both π_i are close to 0 or 1 than when they are not. For a study comparing two treatments on the proportion of subjects who die, the difference between 0.010 and 0.001 is more noteworthy than the difference between 0.410 and 0.401, even though both are 0.009. In such cases, the ratio of proportions is also informative.

The *relative risk* is defined to be the ratio of probabilities,

$$(2.3) \text{ relative risk} = \pi_1/\pi_2.$$

It can be any nonnegative real number. A relative risk of 1.0 corresponds to independence. For the proportions just given, the relative risks are $0.010/0.001 = 10.0$ and $0.410/0.401 = 1.02$. Comparing the rows on the second response category gives a different relative risk, $(1 - \pi_1)/(1 - \pi_2)$.

2.2.3 Odds Ratio

For a probability π of success, the *odds* are defined to be

$$\text{odds } \Omega = \pi/(1 - \pi).$$

The odds are nonnegative, with $\Omega > 1.0$ when a success is more likely than a failure. When $\pi = 0.75$, for instance, then $\Omega = 0.75/0.25 = 3.0$; a success is three times as likely as a failure, and we expect about three successes for every one failure. When $\Omega = \frac{1}{3}$, a failure is three times as likely as a success. Inversely,

$$\pi = \Omega/(\Omega + 1).$$

For instance, when the odds $\Omega = \frac{1}{3}$, then the probability $\pi = 0.25$.

Refer again to a 2×2 table. Within row i , the odds of success instead of failure are $\Omega_i = \pi_i(1 - \pi_i)$.

The ratio of the odds Ω_1 and Ω_2 in the two rows,

$$(2.4) \quad \theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)},$$

is called the *odds ratio*.

For joint distributions with cell probabilities $\{\pi_{ij}\}$, the equivalent definition for the odds in row i is $\Omega_i = \pi_{i1}/\pi_{i2}$, $i = 1, 2$. Then the odds ratio is

$$(2.5) \quad \theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

An alternative name for θ is the *cross-product ratio*, because it equals the ratio of the products $\pi_{11}\pi_{22}$ and $\pi_{12}\pi_{21}$ of probabilities from diagonally opposite cells (Yule 1900, 1912).

2.2.4 Properties of the Odds Ratio

The odds ratio can equal any nonnegative number. The condition $\Omega_1 = \Omega_2$ and hence (when all cell probabilities are positive) $\theta = 1$ corresponds to independence of X and Y . When $1 < \theta < \infty$, subjects in row 1 are more likely to have a success than are subjects in row 2; that is, $\pi_1 > \pi_2$. For instance, when $\theta = 4$, the odds of success in row 1 are four times the odds in row 2. This does not mean that the *probability* $\pi_1 = 4\pi_2$; that is the interpretation of a *relative risk* of 4.0. When $0 < \theta < 1$, then $\pi_1 < \pi_2$.

Values of θ farther from 1.0 in a given direction represent stronger association. Two values represent the same association, but in opposite directions, when one is the reciprocal of the other. For instance, when $\theta = 0.25$, the odds of success in row 1 are 0.25 times the odds in row 2, or equivalently, the odds of success in row 2 are $1/0.25 = 4.0$ times the odds in row 1. When the order of the rows is reversed or the order of the columns is reversed, the new value for θ is the reciprocal of the original value.

For inference, we shall see it is sometimes convenient to use $\log \theta$. Independence corresponds to $\log \theta = 0$. The log odds ratio is symmetric about this value—reversal of rows or of columns results in a change in its sign. Two values for $\log \theta$ that are the same except for sign, such as $\log 4 = 1.39$ and $\log 0.25 = -1.39$, represent the same strength of association.

The odds ratio does not change value when the orientation of the table reverses so that the rows become the columns and the columns become the rows. This is clear from the symmetric form of (2.5). It is unnecessary to identify one classification as the response variable in order to use θ . In fact, although (2.4) defined the odds ratio in terms of odds using $\pi_i = P(Y = 1|X = i)$, we could just as well define it using reverse conditional probabilities. With a joint distribution, conditional distributions exist in each direction, and

$$\begin{aligned}\theta &= \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{P(Y = 1|X = 1)/P(Y = 2|X = 1)}{P(Y = 1|X = 2)/P(Y = 2|X = 2)} \\ (2.6) \quad &= \frac{P(X = 1|Y = 1)/P(X = 2|Y = 1)}{P(X = 1|Y = 2)/P(X = 2|Y = 2)}.\end{aligned}$$

Because of this, the odds ratio is equally valid for prospective, retrospective, or cross-sectional sampling designs. The sample odds ratio estimates the same parameter in each case.

For cell counts $\{n_{ij}\}$, the sample odds ratio is

$$\hat{\theta} = (n_{11}n_{22})/(n_{12}n_{21}).$$

This does not change when both cell counts within any row are multiplied by a nonzero constant or when both cell counts within any column are multiplied by a nonzero constant. An implication is that the sample odds ratio estimates the same characteristic (θ) even when the sample is disproportionately large or small from marginal categories of a variable. For a case-control study of the association between vaccination and catching the flu, the sample odds ratio estimates the same characteristic with a random sample of (1) 100 people who got the flu and 100 people who did not, or (2) 40 people who got the flu and 160 people who did not. The sample versions of the difference of proportions and relative risk (2.3) are invariant to multiplication of counts within rows by a constant, but they change with multiplication within columns or with row-column interchange.

2.2.5 Example: Association Between Heart Attacks and Aspirin Use

We illustrate the three association measures with [Table 2.1](#) on aspirin use and heart attacks. The table differentiates between fatal and nonfatal heart attacks, but we combine these outcomes for now.

Of the 11,034 physicians taking placebo, 189 suffered heart attacks, a proportion of $189/11,034 = 0.0171$. Of the 11,037 taking aspirin, 104 had heart attacks, a proportion of 0.0094. The sample difference of proportions is $0.0171 - 0.0094 = 0.0077$. The sample relative risk is $0.0171/0.0094 = 1.82$. The proportion suffering heart attacks of those taking placebo was 1.82 times the proportion suffering heart attacks of those taking aspirin. The sample odds ratio is $(189 \times 10,933)/(10,845 \times 104) = 1.83$. The odds of heart attack for those taking placebo was 1.83 times the odds for those taking aspirin.

2.2.6 Case–Control Studies and the Odds Ratio

With retrospective sampling designs, such as case–control studies, it is possible to estimate conditional probabilities of form $P(X = i|Y = j)$. It is usually not possible to estimate the probability $P(Y = j|X = i)$ of an outcome of interest or the difference of proportions or relative risk for that outcome. It is possible to estimate the odds ratio, however, since by (2.6) it is determined by conditional probabilities in *either* direction.

To illustrate, we revisit [Table 2.5](#) on $X =$ smoking behavior and $Y =$ lung cancer. The data were two binomial samples on X at fixed levels of Y . Thus, we can estimate the probability a subject was a smoker, given the outcome on whether the subject had lung cancer; this was 688/709 for the cases and 650/709 for the controls. We cannot estimate the probability of lung cancer, given whether one smoked, which is more relevant. Thus, we cannot estimate differences or ratios of probabilities of lung cancer. The difference of proportions and relative risk are limited to comparisons of the probabilities of being a smoker. However, we can compute the odds ratio using the sample analog of (2.6),

$$\frac{(688/709)/(21/709)}{(650/709)/(59/709)} = \frac{688 \times 59}{650 \times 21} = 3.0.$$

Moreover, by (2.6), interpretations can use the direction of interest, even though the study was retrospective: The estimated odds of lung cancer for smokers were 3.0 times the estimated odds for nonsmokers.

2.2.7 Relationship Between Odds Ratio and Relative Risk

From definitions [\(2.3\)](#) and [\(2.4\)](#),

$$\text{odds ratio} = \text{relative risk} \left(\frac{1 - \pi_2}{1 - \pi_1} \right).$$

Their magnitudes are similar whenever the probability π_i , of the outcome of interest is close to zero for both groups. We saw this similarity in Section 2.2.5 for the aspirin study, where the heart attack proportion was less than 0.02 for each group. The relative risk was 1.82 and the odds ratio was 1.83.

Because of this similarity, when each π_i is small, the odds ratio provides a rough indication of the relative risk when it is not directly estimable, such as in case-control studies (Cornfield 1951). For instance, for [Table 2.5](#), if the probability of lung cancer is small regardless of smoking behavior, 3.0 is also a rough estimate of the relative risk; that is, for the way smoking was defined in that study, smokers had about 3.0 times the chance of lung cancer as nonsmokers.

2.3 CONDITIONAL ASSOCIATION IN STRATIFIED 2×2 TABLES

An important part of any observational study is the choice of control variables. In studying the effect of X on Y , we should attempt to adjust or “control” any covariate that can influence that relationship. This involves using some mechanism to hold the covariate constant. Otherwise, an observed effect of X on Y may actually reflect effects of that covariate on both X and Y . The relationship between X and Y then shows *confounding*. Experimental studies can remove effects of confounding covariates by randomly assigning subjects to different levels of X , but this is not possible with observational studies.

Suppose that a study considers effects of passive smoking, the effects on a nonsmoker of living with a smoker. To analyze whether passive smoking is associated with lung cancer, a cross-sectional study might compare lung cancer rates between nonsmokers whose spouses smoke and nonsmokers whose spouses do not smoke. The study should attempt to control for age, socioeconomic status, and other variables that might relate both to spouse smoking and to developing lung cancer. Otherwise, results will have limited usefulness. Spouses of nonsmokers may tend to be younger than spouses of smokers, and younger people are less likely to have lung cancer. Then a lower proportion of lung cancer cases among spouses of nonsmokers may merely reflect their lower average age.

In this section we discuss the analysis of the association between categorical variables X and Y while controlling for a possibly confounding variable Z . For simplicity, the examples refer to a single control variable. In later chapters we treat more general cases and use models to perform statistical control.

2.3.1 Partial Tables

A three-way contingency table cross-classifies X , Y , and Z . We control for Z by studying the XY relationship at fixed levels of Z . Two-way cross-sectional slices of the three-way table cross-classify X and Y at separate categories of Z . These cross sections are called *partial tables*. They display the XY relationship while removing the effect of Z by holding its value constant.

The two-way contingency table obtained by combining the partial tables is called the *XY marginal table*. Each cell count in the marginal table is a sum of counts from the same location in the partial tables. The marginal table, rather than controlling Z , ignores it. The marginal table contains no information about Z . It is simply a two-way table relating X and Y but may reflect the effects of Z on X and Y .

The associations in partial tables are called *conditional associations*, because they refer to the association between X and Y conditional on fixing Z at some level. Conditional associations in partial tables can be quite different from associations in marginal tables. In fact, it can be misleading to analyze only marginal tables of a multiway contingency table. The following example illustrates.

2.3.2 Example: Racial Characteristics and the Death Penalty

[Table 2.6](#) is a $2 \times 2 \times 2$ contingency table—two rows, two columns, and two layers—from an article that studied effects of racial characteristics on whether persons convicted of homicide received the death penalty. The 674 subjects classified in [Table 2.6](#) were the defendants in indictments involving cases with multiple murders in Florida between 1976 and 1987. The variables in [Table 2.6](#) are Y = death penalty verdict, having the categories (yes, no), X = race of defendant, and Z = race of victims, each having the categories (white, black). We study the effect of defendant's race on the death penalty verdict, treating victims' race as a control variable. [Table 2.6](#) has a 2×2 partial table relating defendant's race and the death penalty verdict at each category of victims' race.

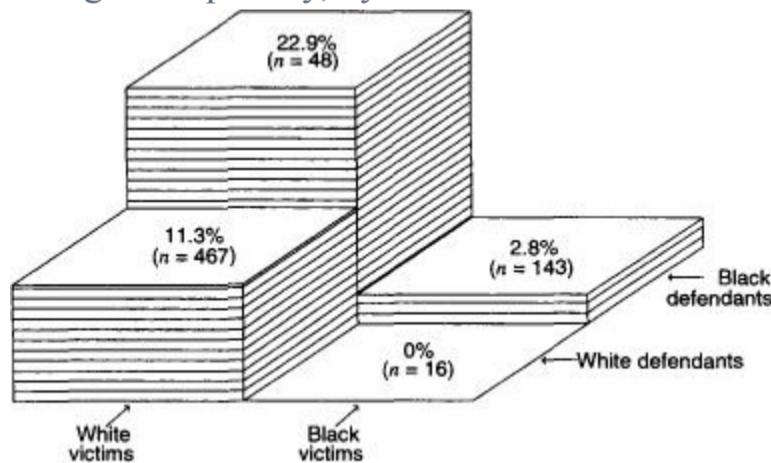
Table 2.6 Death Penalty Verdict by Defendant's Race and Victims' Race

Victims' Race	Defendant's Race	Death Penalty		
		Yes	No	Percent Yes
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8
Total	White	53	430	11.0
	Black	15	176	7.9

Source: M. L. Radelet and G. L. Pierce, *Florida Law Rev.* 43: 1–34, 1991. Reprinted with permission from the *Florida Law Review*.

For each combination of defendant's race and victims' race, [Table 2.6](#) lists and [Figure 2.1](#) displays the percentage of defendants who received the death penalty. These describe the conditional associations. When the victims were white, the death penalty was imposed $22.9\% - 11.3\% = 11.6\%$ more often for black defendants than for white defendants. When the victims were black, the death penalty was imposed 2.8% more often for black defendants than for white defendants. *Controlling* for victims' race by keeping it fixed, the death penalty was imposed more often on black defendants than on white defendants.

Figure 2.1 Percentage receiving death penalty, by defendant's race and victims' race.



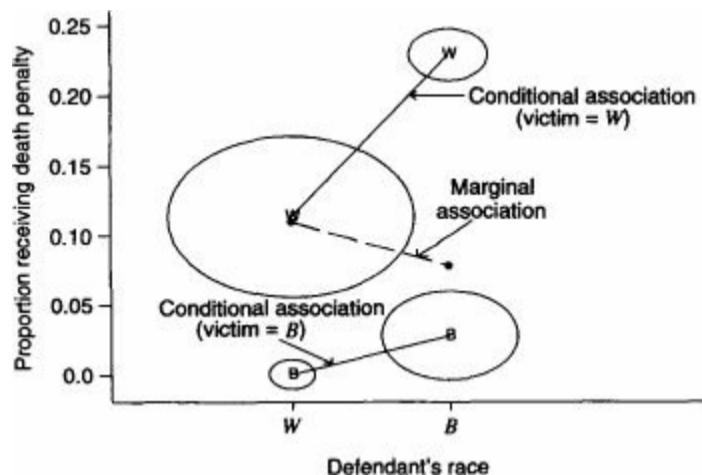
The bottom portion of [Table 2.6](#) displays the marginal table. It results from summing the cell counts in [Table 2.6](#) over the two categories of victims' race, thus combining the two partial tables (e.g., $11 + 4 = 15$). Overall, 11.0% of white defendants and 7.9% of black defendants received the death penalty. *Ignoring* victims' race, the death penalty was imposed less often on black defendants than on white defendants. The association reverses direction compared with the partial tables.

Why does the association change so much when we ignore versus control victims' race? This relates to the nature of the association between victims' race and each of the other variables. First, the association between victims' race and defendant's race is extremely strong. The marginal table relating these variables has odds ratio $(467 \times 143)/(48 \times 16) = 87.0$. Second, [Table 2.6](#) shows that, regardless of defendant's race, the death penalty was much more likely when the victims were white than when the victims were black. So whites are tending to kill whites, and killing whites is more likely to result in the death penalty. This suggests that the marginal association should show a greater

tendency than the conditional associations for white defendants to receive the death penalty. In fact, [Table 2.6](#) has this pattern.

[Figure 2.2](#) illustrates why the marginal association differs so from the conditional associations. For each defendant's race, the figure plots the proportion receiving the death penalty at each category of victims' race. Each proportion is labeled by a letter symbol giving the category of victims' race. Surrounding each observation is a circle having area proportional to the number of observations at that combination of defendant's race and victims' race. For instance, the W in the largest circle represents a proportion of 0.113 receiving the death penalty for cases with white defendants and white victims. That circle is largest because the number of cases at that combination ($53 + 414 = 467$) is largest. The next-largest circle relates to cases in which blacks kill blacks.

Figure 2.2 Proportion receiving death penalty by defendant's race, controlling and ignoring victims' race.



We control for victims' race by comparing circles having the same victims' race letter at their centers. The line connecting the two W circles has a positive slope, as does the line connecting the two B circles. Controlling for victims' race, this reflects the death penalty being more likely for black defendants than for white defendants. When we add results across victims' race to get a summary result for the marginal effect of defendant's race on the death penalty verdict, the larger circles, having the greater number of cases, have greater influence. Thus, the summary proportions for each defendant's race, marked on the figure by periods, fall closer to the center of the larger circles than to the center of the smaller circles. A line connecting the summary marginal proportions has negative slope, indicating that overall the death penalty was more likely for white than for black defendants.

The result that a marginal association can have a different direction from each conditional association is called *Simpson's paradox* (Simpson 1951), although it was noted as early as in Yule (1903). It applies to quantitative as well as categorical variables. Statisticians commonly use it to caution against imputing causal effects from an association of X with Y . For instance, when doctors started to observe association between smoking and lung cancer, statisticians such as R. A. Fisher warned that some variable (e.g., a genetic factor) could exist such that the association would disappear under the relevant control. However, others (e.g., J. Cornfield in 1954, as summarized by Greenhouse 2009) showed that at least as strong an association must exist between a confounding variable Z and both X and Y in order for the effect of X on Y to disappear or change under the control. See Breslow and Day (1980, Sec. 3.4) and Bross (1967) for related comments.

2.3.3 Conditional and Marginal Odds Ratios

Odds ratios can describe marginal and conditional associations. We illustrate for $2 \times 2 \times K$ tables, where K denotes the number of categories of a control variable, Z . Let $\{\mu_{ijk}\}$ denote cell expected frequencies for some sampling model, such as binomial, multinomial, or Poisson sampling.

Within a fixed category k of Z , the odds ratio

$$(2.7) \quad \theta_{XY(k)} = \frac{\mu_{11k} \mu_{22k}}{\mu_{12k} \mu_{21k}}$$

describes conditional XY association in partial table k . The *conditional odds ratios* for the K partial tables can be quite different from the marginal odds ratio. The XY marginal table has expected frequencies $\{\mu_{ij+} = \sum_k \mu_{ijk}\}$. The XY marginal odds ratio is

$$\theta_{XY} = \frac{\mu_{11+} \mu_{22+}}{\mu_{12+} \mu_{21+}}.$$

Sample values of $\theta_{XY(k)}$ and θ_{XY} use similar formulas with cell counts substituted for expected frequencies. We illustrate for the association between defendant's race and the death penalty in [Table 2.6](#). In the first partial table, victims' race is white and

$$\hat{\theta}_{XY(1)} = \frac{53 \times 37}{414 \times 11} = 0.43.$$

The sample odds for white defendants receiving the death penalty were 43% of the sample odds for black defendants. In the second partial table, victims' race is black and the estimated odds ratio equals $\hat{\theta}_{XY(2)} = (0 \times 139)/(16 \times 4) = 0.0$, since the death penalty was never given to white defendants with black victims.

Estimation of the marginal odds ratio uses the 2×2 marginal table within [Table 2.6](#), collapsing over victims' race, or $(53 \times 176)/(430 \times 15) = 1.45$. The sample odds of the death penalty were 45% higher for white defendants than for black defendants. Yet within each victims' race category, those odds were smaller for white defendants. This reversal in the association after controlling for victims' race illustrates Simpson's paradox.

2.3.4 Marginal Independence Versus Conditional Independence

More generally, when X and Y may have multiple categories, an $I \times J \times K$ table describes the relationship between X and Y , controlling for Z . If X and Y are independent in partial table k , then X and Y are said to be *conditionally independent at level k* of Z . When Y is a response, this means that

$$(2.8) P(Y = j|X = i, Z = k) = P(Y = j|Z = k), \text{ for all } i, j.$$

More generally, X and Y are said to be *conditionally independent given Z* when they are conditionally independent at every level of Z , that is, when (2.8) holds for all k . Then, given Z , Y does not depend on X .

Suppose that a single multinomial applies to the entire three-way table, with joint probabilities $\{\pi_{ijk} = P(X = i, Y = j, Z = k)\}$. Then

$$\pi_{ijk} = P(X = i, Z = k) P(Y = j|X = i, Z = k).$$

Under conditional independence of X and Y , given Z , this equals

$$\pi_{i+k} P(Y = j|Z = k) = \pi_{i+k} P(Y = j, Z = k) / P(Z = k).$$

Thus, conditional independence is then equivalent to

$$(2.9) \pi_{ijk} = \pi_{i+k} \pi_{+jk} / \pi_{++k} \text{ for all } i, j, \text{ and } k.$$

Conditional independence does not imply marginal independence (Yule 1903). For instance, summing (2.9) over k on both sides yields

$$\pi_{ij+} = \sum_k (\pi_{i+k} \pi_{+jk} / \pi_{++k}).$$

All three terms in the summation involve k , and this does not simplify to $\pi_{ij+} = \pi_{i++} \pi_{+j+}$, which is marginal independence.

For $2 \times 2 \times K$ tables, X and Y are conditionally independent when the odds ratio between X and Y equals 1.0 at each category of Z . The expected frequencies $\{\mu_{ijk}\}$ in Table 2.7 illustrate this relation for Y = response (success, failure), X = drug treatment (A, B), and Z = clinic (1, 2). From (2.7), the conditional XY odds ratios are

Table 2.7 Expected Frequencies Showing that Conditional Independence Does Not Imply Marginal Independence

Clinic	Treatment	Response	
		Success	Failure
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32
Total	A	20	20
	B	20	40

$$\theta_{XY(1)} = \frac{18 \times 8}{12 \times 12} = 1.0, \quad \theta_{XY(2)} = \frac{2 \times 32}{8 \times 8} = 1.0.$$

Given the clinic, response and treatment are conditionally independent. The marginal table combines the tables for the two clinics. Its odds ratio is $\theta_{XY} = (20 \times 40) / (20 \times 20) = 2.0$, so the variables are not marginally independent.

Ignoring the clinic, why are the odds of a success for treatment A twice those for treatment B? The conditional XZ and YZ odds ratios give a clue. The odds ratio between Z and either X or Y , at each fixed category of the other variable, equals 6.0. For instance, the XZ odds ratio at the first category of Y equals $(18 \times 8) / (12 \times 2) = 6.0$. The conditional odds (given response) of receiving treatment A at clinic 1 are six times those at clinic 2, and the conditional odds (given treatment) of success at clinic 1 are six times those at clinic 2. Clinic 1 tends to use treatment A more often, and clinic 1 also tends to have more successes.

For instance, if patients at clinic 1 tended to be younger and in better health than those at clinic 2, perhaps they had a better success rate regardless of the treatment received.

It is misleading to study only the marginal table, concluding that successes are more likely with

treatment A. Subjects within a particular clinic are likely to be more homogeneous than the overall sample, and response is independent of treatment in each clinic.

2.3.5 Homogeneous Association

A $2 \times 2 \times K$ table has *homogeneous XY association* when

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}.$$

Then the effect of X on Y is the same at each category of Z . Conditional independence of X and Y is the special case in which each $\theta_{XY(k)} = 1.0$.

Under homogeneous XY association, homogeneity also holds for the other associations. For instance, the conditional odds ratio between two categories of X and two categories of Z is identical at each category of Y . For the odds ratio, homogeneous association is a symmetric property. It applies to any pair of variables viewed across the categories of the third. When it occurs, there is said to be *no interaction* between two variables in their effects on the other variable.

When interaction exists, the conditional odds ratio for any pair of variables changes across categories of the third. For $X =$ smoking (yes, no), $Y =$ lung cancer (yes, no), and $Z =$ age (<45 , $45-65$, >65), suppose that $\theta_{XY(1)} = 1.2$, $\theta_{XY(2)} = 3.9$, and $\theta_{XY(3)} = 8.8$. Then smoking has a weak effect on lung cancer for young people, but the effect strengthens considerably with age. Age is called an *effect modifier*; the effect of smoking is modified depending on the value of age.

For the death penalty data ([Table 2.6](#)), $\theta_{XY(1)} = 0.43$ and $\theta_{XY(2)} = 0.0$. The values are not close, but the second estimate is imprecise because of the zero cell count. Because of the ordinary variation that occurs from sampling variability, these partial tables do not necessarily contradict homogeneous association in a population.

Some analyses of categorical data assume homogeneous association, and we'll also see how to test such an assumption. For example, when each 2×2 table results from a particular study, the statistical analysis may combine information from the various studies to summarize the overall evidence against conditional independence and to assess whether the effect was the same in each study. Such an analysis is called a *meta-analysis*. In Section 6.4 we show how to analyze whether sample data are consistent with homogeneous association or conditional independence.

2.3.6 Collapsibility: Identical Conditional and Marginal Associations

Even when conditional associations are identical, we've seen that they may differ from a marginal association. When do they not differ? We'll study this in some detail in Section 10.1, but for now we'll state two basic results, for $2 \times 2 \times K$ tables stratifying by categories of Z :

Collapsibility of Odds Ratios. When $\theta_{XY(k)}$ is identical at every level k of Z , that value equals the marginal odds ratio θ_{XY} if either Z and X are conditionally independent or if Z and Y are conditionally independent.

Collapsibility of Difference of Proportions (or Relative Risk). When $\pi_1 - \pi_2$ (or π_1/π_2) is the same at every level of Z , that value equals the corresponding marginal measure if Z is independent of X in the marginal XZ table or if Z is conditionally independent of Y given X .

The conditions for odds ratio collapsibility state that the variable treated as the control (Z) is conditionally independent of X or Y , or both. For example, the conditional odds ratio between defendant's race and the death penalty verdict is collapsible over victim's race if (1) for each death penalty outcome, victim's race and defendant's race are independent, or (2) for each defendant's race, the chance of the death penalty is the same when the victim was white as when the victim was black. The first condition for collapsibility of the difference of proportions or relative risk is satisfied, for example, for factorial designs with the same number of observations at each combination of levels of X and Z . For details and extensions, see the references in Note 2.3.

2.4 MEASURING ASSOCIATION IN $I \times J$ TABLES

For 2×2 tables, a single number such as the odds ratio can summarize the association. For $I \times J$ tables, it is usually not possible to summarize association by a single number without some loss of information. However, a set of odds ratios or another summary index can describe certain features of the association.

2.4.1 Odds Ratios in $I \times J$ Tables

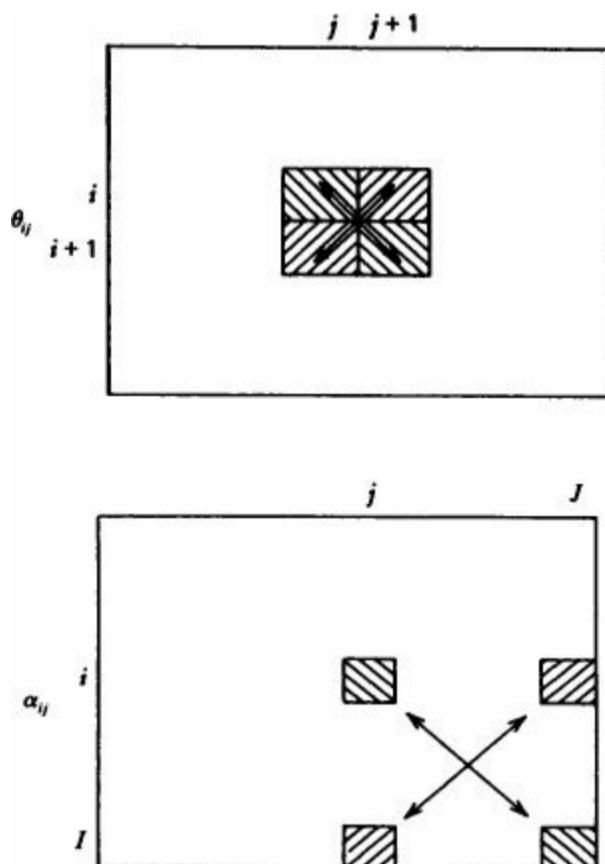
Odds ratios can use each of the $\binom{I}{2}$ pairs of rows in combination with each of the $\binom{J}{2}$ pairs of columns. For rows a and b and columns c and d , the odds ratio $(\pi_{ac}\pi_{bd})/(\pi_{bc}\pi_{ad})$ uses four cells in a rectangular pattern. There are $\binom{I}{2}\binom{J}{2}$ odds ratios of this type. This set of odds ratios contains much redundant information.

Consider the subset of $(I - 1)(J - 1)$ local odds ratios

$$(2.10) \quad \theta_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}}, \quad i = 1, \dots, I - 1, \quad j = 1, \dots, J - 1.$$

[Figure 2.3](#) shows that local odds ratios use cells in adjacent rows and adjacent columns. These $(I - 1)(J - 1)$ odds ratios determine all odds ratios formed from pairs of rows and pairs of columns. To illustrate, in [Table 2.1](#), the sample local odds ratio is 2.08 for the first two columns and 1.74 for the second and third columns. In each case, the more serious outcome was more prevalent for the placebo group. The product of these two odds ratios is 3.63, which is the odds ratio for the first and third columns.

[Figure 2.3](#) Odds ratios for $I \times J$ tables.



Construction (2.10) for a minimal set of odds ratios is not unique. Another basic set is

$$(2.11) \quad \alpha_{ij} = \frac{\pi_{ij}\pi_{IJ}}{\pi_{Ij}\pi_{iJ}}, \quad i = 1, \dots, I - 1, \quad j = 1, \dots, J - 1.$$

This uses the rectangular pattern of cells determined by the cell in row i and column j and the cell in the last row and last column. [Figure 2.3](#) illustrates.

Given the marginal distributions $\{\pi_{i+}\}$ and $\{\pi_{j+}\}$, when $\{\pi_{ij} > 0\}$, conversion of the probabilities into the set of odds ratios (2.10) or (2.11) does not discard information. The cell probabilities determine the odds ratios, and given the marginals, the odds ratios determine the cell probabilities. In this sense, $(I - 1)(J - 1)$ parameters can describe any association in an $I \times J$ table. Independence is equivalent to all $(I - 1)(J - 1)$ odds ratios equaling 1.0.

For three-way $I \times J \times K$ tables, sets of odds ratios in the partial tables describe the conditional association. Homogeneous XY association means that a conditional odds ratio formed using any particular two categories of X and any particular two categories of Y is the same at each category of Z .

2.4.2 Association Factors

An alternative type of association summary focuses on individual cells and whether a cell has more or fewer subjects than we'd expect if the variables are independent. One way to do this uses the *IJ association factors* (Good 1956),

$$\pi_{ij}/(\pi_i + \pi_j).$$

An association factor is the ratio of the cell probability to the probability corresponding to independence for the particular marginal distributions. It falls between 0 and $\min(1/\pi_{i+}, 1/\pi_{+j})$, with the baseline value of 1 corresponding to independence.

It can be informative to investigate which cells have probabilities substantially different from independence. For instance, we could regard the departure from independence in a cell as being noteworthy when the association factor is larger than 2 or smaller than $\frac{1}{2}$.

2.4.3 Summary Measures of Association

Another way to describe association uses a single summary index. We discuss this first for nominal variables and then ordinal variables. The most interpretable indices for nominal variables have the same structure as R -squared for interval variables. It and the more general intraclass correlation coefficient and correlation ratio (Kendall and Stuart 1979) describe the proportional reduction in variance from the marginal distribution of the response Y to the conditional distributions of Y given an explanatory variable X .

Let $V(Y)$ denote a measure of variation for the marginal distribution $\{\pi_{+j}\}$ of Y , and let $V(Y|i)$ denote this measure computed for the conditional distribution $\{\pi_{1|i}, \dots, \pi_{J|i}\}$ of Y at the i th setting of X . A proportional reduction in variation measure has the form

$$(2.12) \quad \frac{V(Y) - E[V(Y|X)]}{V(Y)},$$

where $E[V(Y|X)]$ is the expectation of the conditional variation taken with respect to the distribution of X . For the marginal distribution $\{\pi_{i+}\}$ of X , $E[V(Y|X)] = \sum_i \pi_{i+} V(Y|i)$.

For a nominal response, Theil (1970) proposed an index using the variation measure $V(Y) = \sum_j \pi_{+j} \log \pi_{+j}$, called the *entropy*. For contingency tables, the proportional reduction in entropy equals

$$(2.13) \quad U = -\frac{\sum_i \sum_j \pi_{ij} \log(\pi_{ij}/\pi_{i+} \pi_{+j})}{\sum_j \pi_{+j} \log \pi_{+j}},$$

called the *uncertainty coefficient*. It takes value between 0 and 1: $U = 0$ is equivalent to independence of X and Y ; $U = 1$ is equivalent to a lack of conditional variation, in the sense that for each i , $\pi_{j|i} = 1$ for some j .

Various measures of form (2.12) describe association in $I \times J$ tables (see Exercises 2.39 and 2.40). A difficulty with them is developing intuition for how large a value constitutes a strong association. How do we interpret, say, a 30% reduction in entropy? Summary measures seem easier to interpret and more useful when both classifications are ordinal, as discussed next.

2.4.4 Ordinal Trends: Concordant and Discordant Pairs

[Table 2.8](#) cross-classifies job satisfaction with age for a recent General Social Survey (GSS). The GSS is a probability sample of Americans conducted every other year. Both classifications are ordinal as measured, with the job satisfaction categories being 1 = not satisfied, 2 = fairly satisfied, 3 = very or completely satisfied.

Table 2.8 Cross-Classification of Job Satisfaction by Age of Respondent

Age	Job Satisfaction		
	1	2	3
<30	34	53	88
30–50	80	174	304
>50	29	75	172

Source: 2006 General Social Survey, National Opinion Research Center.

When X and Y are ordinal, a monotone trend association is common. For instance, perhaps job satisfaction tends to increase as age does. Measures that describe the degree to which a relationship is monotone can be based on classifying each pair of subjects as concordant or discordant. A pair is *concordant* if the subject ranked higher on X also ranks higher on Y . The pair is *discordant* if the subject ranking higher on X ranks lower on Y .

For [Table 2.8](#), consider a pair of subjects, one in the cell ($<30, 1$) and the other in the cell ($30–50, 2$). This pair is concordant, since the second subject ranks higher than the first both on age and on job satisfaction. All 34 subjects in cell ($<30, 1$) form concordant pairs when matched with each of the 174 subjects classified ($30–50, 2$), so these two cells provide $34 \times 174 = 5916$ concordant pairs. Each subject in the cell ($<30, 1$) is also part of a concordant pair when matched with each of the other $(80 + 75 + 29)$ subjects ranked higher on both variables. Similarly, the 53 subjects in the ($<30, 2$) cell are part of concordant pairs when matched with the $(80 + 29)$ subjects ranked higher on both variables. The total number of concordant pairs, denoted by C , equals

$$C = 34(174 + 304 + 75 + 172)$$

$$+ 53(304 + 172) + 80(75 + 172) + 174(172) = 99,566.$$

The total number of discordant pairs of observations is

$$D = 88(80 + 174 + 29 + 75) + 53(80 + 29) + 304(29 + 75) + 174(29) = 73,943.$$

In this example, $C > D$, suggesting a tendency for higher age to be associated with higher job satisfaction.

Consider two independent observations from a joint probability distribution $\{\pi_{ij}\}$. For that pair, the probabilities of concordance and discordance are

$$\prod_c = 2 \sum_i \sum_j \pi_{ij} \left(\sum_{h>i} \sum_{k>j} \pi_{hk} \right), \quad \prod_d = 2 \sum_i \sum_j \pi_{ij} \left(\sum_{h>i} \sum_{k<j} \pi_{hk} \right).$$

Here i and j are fixed in the inner summations, and the factor of 2 occurs because the first observation could be in cell (i, j) and the second in cell (h, k) , or vice versa.

2.4.5 Ordinal Measure of Association: Gamma

Given that a pair is untied on both variables, $\Pi_c / (\Pi_c + \Pi_d)$ is the probability of concordance and $\Pi_d / (\Pi_c + \Pi_d)$ is the probability of discordance. The difference between these probabilities,

$$(2.14) \quad \gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d},$$

is called *gamma* (Goodman and Kruskal 1954). The sample version is $\hat{\gamma} = (C - D)/(C + D)$.

For [Table 2.8](#), $C = 99,566$ and $D = 73,943$. Hence,

$$\hat{\gamma} = (99,566 - 73,943)/(99,566 + 73,943) = 0.148.$$

Only a weak tendency exists for job satisfaction to increase as age increases. Of the untied pairs, the proportion of concordant pairs is 0.148 higher than the proportion of discordant pairs.

Like the correlation, gamma treats the variables symmetrically and it has range $-1 \leq \gamma \leq 1$. A reversal in the category orderings of one variable causes a change in the sign of γ . Whereas the absolute value of the correlation is 1 when the relationship between X and Y is perfectly linear, only monotonicity is required for $|\gamma| = 1$, with $\gamma = 1$ if $\Pi_d = 0$ and $\gamma = -1$ if $\Pi_c = 0$. Independence implies that $\gamma = 0$, but the converse is not true. For instance, a U-shaped joint distribution can have $\Pi_c = \Pi_d$ and hence $\gamma = 0$.

For continuous variables, samples can be fully ranked; that is, no ties occur. Then, $C + D = \binom{n}{2}$ and $\hat{\gamma} = (C - D)/\binom{n}{2}$. This is *Kendall's tau*.

2.4.6 Probabilistic Comparisons of Two Ordinal Distributions

Now consider the special case of a $2 \times J$ table, for comparing two groups on an ordinal response variable Y . Let Y_1 and Y_2 denote the column numbers of the response variable for subjects selected at random from rows 1 and 2, independently of each other. A measure that summarizes their relative size is

$$(2.15) \Delta = P(Y_1 > Y_2) - P(Y_2 > Y_1).$$

Related useful measures are $P(Y_1 > Y_2) + (\frac{1}{2})P(Y_1 = Y_2)$ (Exercise 2.41) and $P(Y_1 > Y_2)/P(Y_2 > Y_1)$ (Agresti 2010, Sec. 2.1.4).

If Y_1 and Y_2 are identically distributed, then $\Delta = 0$. When $\Delta > 0 (< 0)$, then outcomes of Y_1 tend to be larger (smaller) than outcomes of Y_2 . Let $F_{j|i} = \pi_{1|i} + \dots + \pi_{j|i}$. When $F_{j|1} \leq F_{j|2}$ for $j = 1, \dots, J$, the conditional distribution in row 1 is *stochastically higher* than the one in row 2. This condition implies that $\Delta \geq 0.0$.

With sample data in the form of two independent multinomials, we can estimate Δ by

$$\hat{\Delta} = \sum_{j>k} p_{j|1} p_{k|2} - \sum_{j<k} p_{j|1} p_{k|2}.$$

If we artificially identify row 1 as the higher level of the group variable, then this relates to the numbers of concordant and discordant pairs by

$$\hat{\Delta} = (C - D)/(n_1 n_2).$$

With $J = 2$ the measure simplifies to the difference of proportions.

2.4.7 Example: Comparing Pain Ratings After Surgery

[Table 2.9](#) is from a study to compare an active treatment with a control treatment for patients having shoulder tip pain after laparoscopic surgery. The two treatments were randomly assigned to 41 patients. The patients rated their pain level on a scale from 1 (low) to 5 (high) on the fifth day after the surgery.

Table 2.9 Shoulder Tip Pain Scores After Laparoscopic Surgery

Treatments	Pain Scores					Total
	1	2	3	4	5	
Active	19	2	1	0	0	22
Control	7	3	4	3	2	19

Source: T. Lumley, *Biometrics* 52: 354–361, 1996.

The sample conditional distributions on shoulder tip pain are:

Active: (0.86, 0.09, 0.05, 0.00, 0.00)

Control: (0.37, 0.16, 0.21, 0.16, 0.11).

The groups are stochastically ordered, with active treatment patients tending to be lower in their pain rating. For these data,

$$\hat{\Delta} = \frac{[1(7+3) + 2(7)] - [19(3+4+3+2) + 2(4+3+2) + 1(3+2)]}{22 \times 19} = -0.543$$

estimates the difference between the probability that the pain rating is higher for active than control treatments and the probability that the pain rating is higher for control than active treatments.

2.4.8 Correlation for Underlying Normality

For ordinal variables, another approach to measuring association uses the correlation. In simplest form, you merely assign fixed scores or midrank scores to the rows and to the columns and use the ordinary Pearson correlation formula.

An alternative approach, advocated by Karl Pearson, estimates the correlation for a bivariate normal distribution assumed to underlie the contingency table. Pearson (1904) applied this approach for 2×2 tables, where his *tetrachoric correlation* is the ML estimate of the correlation for the bivariate normal. This is the correlation value in the bivariate normal density that produces cell probabilities equal to the sample cell proportions when that density is collapsed to a 2×2 table having the same marginal proportions as the observed table. This approach was later generalized to a *polychoric correlation* for $I \times J$ tables (Tallis 1962).

As Section 17.1 discusses, a strong disagreement arose between Pearson and others about when it was sensible to assume underlying normality for inherently categorical variables. Pearson considered approximating underlying normal correlations in various ways. For example, his *contingency coefficient* (Pearson 1904, Exercise 3.32) is a function of a chi-squared statistic for $I \times J$ tables, and his *biserial correlation* (Pearson 1909) applies to $2 \times c$ tables with ordered columns.

NOTES

Section 2.2: Comparing Two Proportions

2.1 Odds ratio invariance: Breslow (1996) reviewed the development of methods for case-control studies. For 2×2 tables, Edwards (1963) showed that functions of the odds ratio are the only statistics that are invariant both to row–column interchange and to multiplication within rows or within columns by a constant. For $I \times J$ tables, Altham (1970) gave related results. Yule (1912, p. 587) had argued that multiplicative invariance is a desirable property for measures of association, especially when proportions sampled in various marginal categories are arbitrary. Goodman (2000) showed five ways of viewing association in a 2×2 table and proposed a general measure that includes all five.

Section 2.3: Conditional Association in Stratified 2×2 Tables

2.2 Simpson's paradox: Paik (1985) proposed circle diagrams of type [Figure 2.2](#) to summarize three-way tables. For more on Simpson's paradox and when it can happen, see Blyth (1972), Davis (1989), Dong (2005), Greenland et al. (1999), Pavlides and Perlman (2009), Samuels (1993), and Simpson(1951). Good and Mittal (1987) extended it to an *amalgamation paradox*, whereby a marginal measure is greater than the maximum or less than the minimum of the partial table measures.

2.3 Collapsibility: For $I \times J \times 2$ tables, the odds ratio collapsibility conditions in Section 2.3.6 are necessary as well as sufficient (Simpson 1951, Whittemore 1978). For $I \times J \times K$ tables, Ducharme and Lepage (1986) showed the conditions are necessary and sufficient for the odds ratios to remain the same no matter how the levels of Z are pooled (i.e., no matter how Z is partially collapsed). For collapsibility for the difference of proportions and relative risk, see Geng (1992), Shapiro (1982), and Wermuth (1987).

Section 2.4: Measuring Association in $I \times J$ Tables

2.4 Surveys: Goodman and Kruskal (1954, 1959) surveyed the historical development of measures of association and introduced new measures. Agresti (2010, Chaps. 2 and 7) and Kruskal (1958) surveyed ordinal measures of association.

EXERCISES

Applications

2.1 According to the FBI website (www.fbi.gov), in 2008, of female murder victims, 1710 were slain by males and 200 by females, whereas of male murder victims, 4351 were slain by males and 455 by females. Let Y denote sex of victim and X denote sex of offender. Report the sample (a) joint distribution of X and Y , (b) conditional distribution of Y given X , and (c) conditional distribution of X given Y .

2.2 According to the FBI website, of all blacks slain in 2008, 92% were slain by blacks, and of all whites slain in 2005, 85% were slain by whites. Let Y denote race of victim and X denote race of offender.

- Which conditional distribution do these statistics refer to, Y given X , or X given Y ?
- Given that a murderer was white, what additional information would you need to estimate the probability that the victim was white? [Hint: How could you use Bayes' theorem?]
- Consider the previous exercise. Which association is stronger—between sex of victim and sex of offender, or between race of victim and race of offender? Justify your answer.

2.3 An article in *The New York Times* (Feb. 17, 1999) about the PSA blood test for detecting prostate cancer stated: “The test fails to detect prostate cancer in 1 in 4 men who have the disease (false-negative results), and as many as two-thirds of the men tested receive false-positive results.” Let $C(\bar{C})$ denote the event of having (not having) prostate cancer, and let $+(-)$ denote a positive (negative) test result. Which is true: $P(-|C) = \frac{1}{4}$ or $P(C|-) = \frac{1}{4}$? $P(\bar{C}|+) = \frac{2}{3}$ or $P(+|\bar{C}) = \frac{2}{3}$? Determine the sensitivity and specificity.

2.4 [Table 2.10](#) shows fatality results for drivers and passengers in auto accidents in Florida in 2008, according to whether the person was wearing a seat belt.

Table 2.10 Data for Exercise 2.4 on Auto Accidents

Seat-Belt Use	Injury	
	Fatal	Nonfatal
No	1085	55,623
Yes	703	441,239

Source: Florida Department of Highway Safety and Motor Vehicles, www.flhsmv.gov/hsmvdocs/CS2008.pdf.

- Estimate the probability of fatality, conditional on seat-belt use in category (i) no and (ii) yes.
- Estimate the probability of wearing a seat belt, conditional on the injury being (i) fatal and (ii) nonfatal.
- For the most natural choice of response variable, find and interpret the difference of proportions, relative risk, and odds ratio. Why are the relative risk and odds ratio approximately equal?

2.5 Consider the following two studies reported in *The New York Times*.

- A British study reported (Dec. 3, 1998) that of smokers who get lung cancer, “women were 1.7 times more vulnerable than men to get small-cell lung cancer.” Is 1.7 the odds ratio or the relative risk?
- A National Cancer Institute study about tamoxifen and breast cancer reported (Apr. 7, 1998) that the women taking the drug were 45% less likely to experience invasive breast cancer than were women taking placebo. Find the relative risk for (i) those taking the drug compared with those taking placebo, and (ii) those taking placebo compared with those taking the drug.

2.6 According to a report by the United Nations Office on Drugs and Crime, the number of homicides involving firearms per million people is about 62.4 in the United States, 6.0 in Canada, 5.6 in Australia, and 1.3 in the UK. Use the relative risk to compare the United States

with the other countries. For such data, explain why the relative risk is more informative than the difference of proportions.

2.7 An article in *The Economist* (July 3, 2010) stated that the number of people in prison is 154 per 100,000 in England and Wales, 96 per 100,000 in France, 87 per 100,000 in Germany, and 753 per 100,000 in the United States. Explain how to use the relative risk to compare the U.S. rate to the others.

2.8 At the start of the 2010 World Cup, the betting exchange Betfair stated that the odds against being the winning team were 9/2 for Spain, 11/2 for Brazil, 6/1 for England, and 90/1 for the United States. Find the corresponding prior probabilities of winning for these four teams.

2.9 In a recent survey of people aged 50–71 in the United States summarized by N. Freedman et al. (*Lancet Oncol.* 9: 649–656, 2008), during a follow-up period the annual probability of lung cancer occurrence was about 0.00023 for people who had never smoked and about 0.01284 for current smokers who smoked more than two packs per day. Find and interpret the difference of proportions and the relative risk. Which measure is more informative for these data? Why?

2.10 For adults who sailed on the *Titanic* on its fateful voyage, the odds ratio between gender (female, male) and survival (yes, no) was 11.4. (For data, see R. J. M. Dawson, *J. Statist. Ed.* 3, 1995.)

- a. What is wrong with the interpretation, “The probability of survival for females was 11.4 times that for males”? Give the correct interpretation. When would the quoted interpretation be approximately correct?
- b. The odds of survival for females equaled 2.9. For each gender, find the proportion who survived.

2.11 A research study estimated that under a certain condition, the probability that a subject would be referred for heart catheterization was 0.906 for whites and 0.847 for blacks.

- a. A press release about the study stated that the odds of referral for cardiac catheterization for blacks are 60% of the odds for whites. Explain how they obtained 60% (more accurately, 57%).
- b. An Associated Press story later described the study and said “Doctors were only 60% as likely to order cardiac catheterization for blacks as for whites.” Explain what is wrong with this interpretation. Give the correct percentage for this interpretation.

2.12 A 20-year cohort study of British male physicians (R. Doll and R. Peto, *Br. Med. J.* 2: 1525–1536, 1976) noted that the proportion per year who died from lung cancer was 0.00140 for cigarette smokers and 0.00010 for nonsmokers. The proportion who died from coronary heart disease was 0.00669 for smokers and 0.00413 for nonsmokers.

- a. Describe the association of smoking with each of lung cancer and heart disease, using the difference of proportions, relative risk, and odds ratio. Interpret.
- b. Which response is more strongly related to cigarette smoking, in terms of the reduction in number of deaths that would occur with elimination of cigarettes? Explain.

2.13 For the Women’s Health Study, heart attacks were reported for 198 of 19,934 taking aspirin and for 193 of 19,942 taking placebo (*J. Am. Med. Assoc.* 295: 306–313, 2006). Construct the 2 × 2 table that cross-classifies the treatment with whether a heart attack was reported. Estimate the odds ratio. Interpret. (As of 2006, results suggested that, for women, aspirin was helpful for reducing risk of stroke but not necessarily risk of heart attack.)

2.14 According to poll results released by the Pew Research Center (www.people-press.org) in 2010, when adults in the United States were asked whether there is solid evidence that the average temperature on earth has been getting warmer over the past few decades, the estimated odds of a yes response for a Democrat was 2.96 times higher than for an Independent, and it was 2.08 times higher for an Independent than for a Republican. Find the estimated odds ratio between opinion on global warming and whether one is a Democrat or a Republican. Interpret.

2.15 [Table 2.11](#) refers to applicants to graduate school at the University of California at

Berkeley, for fall 1973. It presents admissions decisions by gender of applicant for the six largest graduate departments. Denote the three variables by A = whether admitted, G = gender, and D = department. Find the sample AG conditional odds ratios and the marginal odds ratio. Interpret, and explain why they give such different indications of the AG association.

Table 2.11 Data for Exercise 2.15 on Graduate Admissions

Department	Whether Admitted			
	Male		Female	
	Yes	No	Yes	No
A	512	313	89	19
B	353	207	17	8
C	120	205	202	391
D	138	279	131	244
E	53	138	94	299
F	22	351	24	317
Total	1198	1493	557	1278

Source: Data from P. Bickel et al., *Science* 187: 398–403, 1975.

2.16 State three “real-world” variables X , Y , and Z for which you expect a marginal association between X and Y but conditional independence controlling for Z .

2.17 Based on murder rates in the United States, an Associated Press story reported that the probability that a newborn child has of eventually being a murder victim is 0.0263 for nonwhite males, 0.0049 for white males, 0.0072 for nonwhite females, and 0.0023 for white females.

- a. Find the conditional odds ratios between race and whether a murder victim, given gender. Interpret. Do these variables exhibit homogeneous association?
- b. Half the newborns are of each gender, for each race. Find the marginal odds ratio between race and whether a murder victim.

2.18 At each age level, the death rate is higher in South Carolina than in Maine, but overall, the death rate is higher in Maine. Explain how this could be possible. [For data, see H. Wainer, *Chance* 12(2): 44, 1999.]

2.19 A study of the death penalty for cases in Kentucky between 1976 and 1991 (T. Keil and G. Vito, *Am. J. Criminal Justice* 20: 17–36, 1995) indicated that the defendant received the death penalty in 8% of the 391 cases in which a white killed a white, in 2% of the 108 cases in which a black killed a black, in 12% of the 57 cases in which a black killed a white, and in 0% of the 18 cases in which a white killed a black. Form the three-way contingency table, obtain the conditional odds ratios between the defendant’s race and the death penalty verdict, interpret those associations, study whether Simpson’s paradox occurs, and explain why the marginal association is so different from the conditional associations.

2.20 [Table 2.12](#) is from an early study on the death penalty in Florida. Analyze these data and show that Simpson’s paradox occurs.

Table 2.12 Data for Exercise 2.20 on the Death Penalty

Victim’s Race	Defendant’s Race	Death Penalty	
		Yes	No
White	White	19	132
	Black	11	52
Black	White	0	9
	Black	6	97

Source: Reprinted with permission from M. L. Radelet, *Am. Sociol. Rev.* 46: 918–927, 1981.

2.21 Smith and Jones are baseball players. Smith has a higher batting average than Jones in each of K years. Is it possible that for the combined data from the K years, Jones has the higher batting average? Explain, creating some data with $K = 2$ to illustrate.

2.22 [Table 2.13](#) summarizes responses from a General Social Survey about homosexual sex and premarital sex. Find and interpret a measure of association.

Table 2.13 Data for Exercise 2.22 on Sexual Attitudes

Premarital Sex	Homosexual Sex			
	Always Wrong	Almost Always Wrong	Wrong Only Sometimes	Not Wrong At All
Always Wrong	300	4	4	17
Almost Always Wrong	78	15	3	14
Wrong Only Sometimes	107	16	46	54
Not Wrong At All	234	32	35	336

Source: General Social Survey, 2008.

2.23 For the data in [Table 2.13](#), the two marginal distributions are dependent rather than independent samples, but the measure Δ can still compare those distributions. Find it, and interpret.

2.24 [Table 2.14](#) cross-classifies job satisfaction by race. Determine whether the groups are stochastically ordered, and estimate the difference between the probability that job satisfaction is higher for blacks than whites and the probability that job satisfaction is higher for whites than blacks.

Table 2.14 Cross-Classification of Job Satisfaction by Race of Respondent

Race	Job Satisfaction			
	Dissatisfied	Neutral	Fairly Satisfied	Very or Completely Satisfied
Black	19	13	42	59
White	47	40	215	430

Source: 2006 General Social Survey, National Opinion Research Center.

Theory and Methods

2.25 For a diagnostic test of a certain disease, let π_1 denote the probability that the diagnosis is positive given that a subject has the disease, and let π_2 denote the probability that the diagnosis is positive given that a subject does not have it. Let ρ denote the probability that a subject has the disease.

a. More relevant to a patient who has received a positive diagnosis is the probability that he or she truly has the disease. Given that a diagnosis is positive, show that the probability that a subject has the disease (called the *positive predictive value*) is

$$\pi_1 \rho / [\pi_1 \rho + \pi_2(1 - \rho)].$$

b. Suppose that a diagnostic test for HIV+ status has both sensitivity and specificity equal to 0.95, and $\rho = 0.005$. Find the probability that a subject is truly HIV+, given that the diagnostic test is positive.

c. To better understand the answer in (b), using the probabilities given there either (i) find the joint probabilities relating diagnosis to actual disease status and discuss their relative sizes, or (ii) construct a tree diagram showing what you would expect to happen for a typical sample of 1000 subjects (first branching from the root according to whether a subject is truly HIV+ and then branching according to the test result), showing that of the subjects with a positive diagnosis, the proportion actually HIV+ agrees with the result in (b).

d. Discuss how the answer in (b) depends on the prevalence ρ . Illustrate by finding the answer when $\rho = 0.10$ instead of 0.005.

2.26 Show that the odds ratio and relative risk need *not* be similar when π_i , is close to 1.0 for both groups.

2.27 Let D denote having a certain disease and E denote having exposure to a certain risk factor. The *attributable risk* (AR) is the proportion of disease cases attributable to that exposure (see Benichou 2005).

a. Let $P(\bar{E}) = 1 - P(E)$. Explain why

$$AR = [P(D) - P(D|\bar{E})]/P(D).$$

b. Show that AR relates to the relative risk RR by

$$AR = [P(E)(RR - 1)]/[1 + P(E)(RR - 1)].$$

2.28 In comparing new and standard treatments with success probabilities π_1 and π_2 , the *number*

needed to treat (NNT) is the number of patients that would need to be treated with the new treatment instead of the standard in order for one patient to benefit. Explain why a natural estimate of this is $1/(\hat{\pi}_1 - \hat{\pi}_2)$.

2.29 For a 2×2 table of counts $\{n_{ij}\}$, show that the odds ratio is invariant to **(a)** inter-changing rows with columns, and **(b)** multiplication of cell counts within rows or within columns by $c \neq 0$. Show that the difference of proportions and the relative risk do not have these properties.

2.30 For given π_1 and π_2 , show that the relative risk cannot be farther than the odds ratio from their independence value of 1.0.

2.31 Let $\pi_{ij|k} = P(X=i, Y=j|Z=k)$. Explain why XY conditional independence is

$$\pi_{ij|k} = \pi_{i+k} \pi_{j+k} \quad \text{for all } i \text{ and } j \text{ and } k.$$

2.32 For a $2 \times 2 \times 2$ table, show that homogeneous association is a symmetric property, by showing that equal XY conditional odds ratios is equivalent to equal YZ conditional odds ratios.

2.33 For a $2 \times 2 \times 2$ table, suppose $\theta_{XY(1)} = \theta_{XY(2)} = \theta$. For a possibly confounding variable Z , let θ_c denote the common value of $\theta_{(i)YZ}$. Let $\pi_1 = P(Z=1|X=1, Y=2)$ and $\pi_2 = P(Z=1|X=2, Y=2)$.

a. Show (Breslow and Day 1980, p. 96) that

$$\theta_{XY} = \theta \frac{\theta_c \pi_1 + (1 - \pi_1)}{\theta_c \pi_2 + (1 - \pi_2)}.$$

b. Verify that either odds ratio collapsibility condition in Section 2.3.6 implies that the *confounding risk ratio* θ_{XY}/θ equals 1.0.

c. Describe what needs to happen for θ_{XY}/θ to be far from 1.0. Illustrate with particular values of $\theta_c > 1$ and $\pi_1 > \pi_2$. Describe a study in which such values would be plausible.

2.34 When X and Y are conditionally dependent at each level of Z yet marginally independent, Z is called a *suppressor variable*. Specify joint probabilities for a $2 \times 2 \times 2$ table to show that this can happen **(a)** when there is homogeneous association, and **(b)** when the association has opposite direction in the partial tables.

2.35 Show that the $\{\alpha_{ij}\}$ in (2.11) determine all odds ratios formed from pairs of rows and pairs of columns.

2.36 For $I \times J$ contingency tables, explain why the variables are independent when the $(I-1)(J-1)$ differences $\pi_{j|I} - \pi_{j|I} = 0$, $i = 1, \dots, I-1, j = 1, \dots, J-1$.

2.37 Suppose that $\{Y_{ij}\}$ are independent Poisson variates with means $\{\mu_{ij}\}$. Show that $P(Y_{ij} = n_{ij})$ for all i, j , conditional on $[Y_{i+} = n_i]$, satisfy independent multinomial sampling [i.e., the product of (2.2) for all i] within the rows.

2.38 For 2×2 tables. Yule (1900, 1912) introduced

$$Q = \frac{\pi_{11} \pi_{22} - \pi_{12} \pi_{21}}{\pi_{11} \pi_{22} + \pi_{12} \pi_{21}},$$

which he labeled Q in honor of the Belgian statistician Quetelet. It is now called *Yule's Q*.

a. Show that for 2×2 tables, Goodman and Kruskal's $\gamma = Q$.

b. Show that Q relates to the odds ratio by $Q = (\theta - 1)/(\theta + 1)$, a monotone transformation of θ from the $[0, \infty]$ scale onto the $[-1, +1]$ scale.

2.39 Goodman and Kruskal (1954) proposed an association measure (τ) for nominal variables based on variation measure

$$V(Y) = \sum \pi_{+j}(1 - \pi_{+j}) = 1 - \sum \pi_{+j}^2.$$

a. Show that $V(Y)$ is the probability that two independent observations on Y fall in different categories. Show that $V(Y) = 0$ when $\pi_{+j} = 1$ for some j and $V(Y)$ takes maximum value of $(J-1)/J$ when $\pi_{+j} = 1/J$ for all j . This index relates to measures of *concentration* and *diversity* proposed for various applications, such as by Corrado Gini (1914a), who was highly influential in the twentieth century in the development of descriptive statistics in Italy, and by E. H. Simpson (1949) who described species diversity (see Exercise 16.13).

b. For the proportional reduction in variation, show that $E[V(Y|X)] = 1 - \sum_i \sum_j \pi_{ij}^2 / \pi_{i+}$. [The resulting measure [\(2.12\)](#) is called the *concentration coefficient*. Like the uncertainty coefficient U , $\tau = 0$ is equivalent to independence. Haberman (1982) presented generalized concentration and uncertainty coefficients.]

2.40 The measure of association *lambda* for nominal variables (Goodman and Kruskal 1954) has $V(Y) = 1 - \max\{\pi_{+j}\}$ and $V(Y|i) = 1 - \max_j \{\pi_{j|i}\}$. Interpret lambda as a proportional reduction in error for predictions which select the response category that is most likely. Show that independence implies $\lambda = 0$ but that the converse is not true.

2.41 Show that Δ in [\(2.15\)](#) relates to $\alpha = P(Y_1 > Y_2) + (\frac{1}{2})P(Y_1 = Y_2)$ by

$$\alpha = (\Delta + 1)/2, \quad \Delta = 2\alpha - 1,$$

with α having range $[0, 1]$ and null value $\frac{1}{2}$.

CHAPTER 3

Inference for Two-Way Contingency Tables

In this chapter we introduce inferential methods for contingency tables. Many of these methods also play a vital role in analyses, presented in later chapters, for which categorical data need not have contingency table form—such as when some explanatory variables are continuous. The methods assume a standard sampling scheme for categorical data—Poisson, multinomial, or independent multinomial (or binomial) sampling.

In Section 3.1 we present confidence intervals for measures of association, such as the odds ratio and the difference and ratio of proportions. Section 3.2 introduces chi-squared tests of the hypothesis of independence between two categorical variables and confidence intervals obtained by inverting more general chi-squared tests. In Section 3.3 we show how to follow-up chi-squared tests using residuals and the partitioning property of chi-squared to extract components that describe the evidence about the association. For ordinal variables, in Section 3.4 we present more powerful inference that utilizes the category orderings. The methods of Sections 3.1 through 3.4 assume large samples. In Section 3.5 we introduce small-sample methods. In Section 3.6 we present Bayesian methods of inference for contingency tables.

3.1 CONFIDENCE INTERVALS FOR ASSOCIATION PARAMETERS

The precision of estimators of association parameters is characterized by standard errors of their sampling distributions. In this section we present standard errors and simple confidence intervals, focusing on parameters for 2×2 tables. We'll present alternative intervals, based on inverting score and likelihood-ratio tests, in Sections 3.2.5 and 3.2.6.

3.1.1 Interval Estimation of the Odds Ratio

The sample odds ratio for a 2×2 table is $\hat{\theta} = (n_{11}n_{22})/(n_{12}n_{21})$. For a multinomial sample, the estimator $\hat{\theta}$ has an asymptotic normal distribution around θ . Unless n is very large, however, its sampling distribution is highly skewed. When $\theta = 1$, for instance, $\hat{\theta}$ cannot be much smaller than θ (since $\hat{\theta} \geq 0$), but it could be much larger with nonnegligible probability. The log transform, having an additive rather than multiplicative structure, converges more rapidly to normality. An estimated standard error for $\log \hat{\theta}$ is

$$(3.1) \quad \hat{\sigma}(\log \hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

We derive this formula in Section 3.1.7.

By the large-sample normality of $\log \hat{\theta}$,

$$(3.2) \quad \log \hat{\theta} \pm z_{\alpha/2} \hat{\sigma}(\log \hat{\theta})$$

is a Wald confidence interval for $\log \theta$ (Woolf 1955). Exponentiating (taking antilogs of) its endpoints provides a confidence interval for θ . The actual coverage probability is usually a bit higher than the nominal level.

If an $n_{ij} = 0$, $\hat{\theta}$ equals 0 or ∞ and the Wald interval does not exist. Since such an outcome has positive probability, the actual expected value and variance of $\hat{\theta}$ and $\log \hat{\theta}$ do not exist¹. This is not problematic for confidence intervals formed by inverting the score test or likelihood-ratio test for θ . For these intervals, when $\hat{\theta} = 0$, 0 is the lower limit and when $\hat{\theta} = \infty$, ∞ is the upper limit. This is sensible for a frequentist approach. This also happens when we construct a small-sample confidence interval for the odds ratio to be introduced in Section 16.6.4. Alternatively, but somewhat ad hoc, we can use the Wald formula (3.2) following some adjustment, such as by replacing $\{n_{ij}\}$ by $\{n_{ij} + 0.5\}$ in the estimator and standard error. In terms of bias and mean squared error, Gart and Zweifel (1967) and Haldane (1956) showed that such amended estimators perform well (see also Exercise 16.8).

3.1.2 Example: Seat-Belt Use and Traffic Deaths

We illustrate inference for the odds ratio with [Table 3.1](#), which shows fatality results for children under age 18 who were passengers in auto accidents in Florida in 2008, according to whether the child was wearing a seat belt. The sample odds ratio $\hat{\theta} = 10.83$, and the standard error [\(3.1\)](#) of $\log \hat{\theta} = 2.383$ is $\hat{\sigma}(\log \hat{\theta}) = 0.242$. A 95% confidence interval for $\log \theta$ in the population this sample represents is $2.383 \pm 1.96(0.242)$, or $(1.908, 2.857)$.

Table 3.1 Injury Outcome and Seat-Belt Use for Child Passengers in Automobile Accidents in Florida in 2008

Seat-Belt Use	Injury Outcome		
	Fatal	Nonfatal	Total
No	54	10,325	10,379
Yes	25	51,790	51,815

Source: Florida Department of Highway Safety and Motor Vehicles,
www.flhsmv.gov/hsmvdocs/CS2008.pdf.

The corresponding interval for θ is $[\exp(1.908), \exp(2.857)]$ or $(6.74, 17.42)$. There is a very strong association. Even though the overall sample size is extremely large, the estimate of the true odds ratio is rather imprecise because of the relatively small number of fatalities (Exercise 3.25).

3.1.3 Interval Estimation of Difference of Proportions and Relative Risk

The difference of proportions and the relative risk compare conditional distributions of a response variable for two groups. For these measures, we treat the samples as independent binomials. For group i , Y_i has a binomial distribution with sample size n_i and a probability π_i of a “success” outcome.

The sample proportion $\hat{\pi}_i = y_i/n_i$ has expectation π_i and variance $\pi_i(1 - \pi_i)/n_i$. Since $\hat{\pi}_1$ and $\hat{\pi}_2$ are independent, their difference has $E(\hat{\pi}_1 - \hat{\pi}_2) = \pi_1 - \pi_2$ and standard error

$$(3.3) \quad \sigma(\hat{\pi}_1 - \hat{\pi}_2) = \sqrt{\frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}}.$$

The estimate $\hat{\sigma}(\hat{\pi}_1 - \hat{\pi}_2)$ replaces π_i by $\hat{\pi}_i$. Then

$$(3.4) \quad (\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\alpha/2} \hat{\sigma}(\hat{\pi}_1 - \hat{\pi}_2)$$

is a Wald confidence interval for $\pi_1 - \pi_2$. Like the Wald interval (1.13) for a single proportion, it usually has true coverage probability less than the nominal confidence level, especially when π_1 and π_2 are near 0 or 1. Section 3.2.5, Note 3.1, and Exercise 3.27 present other methods.

The sample relative risk is $r = \hat{\pi}_1/\hat{\pi}_2 = [(y_1/n_1)/(y_2/n_2)]$. Like the odds ratio, it converges to normality faster on the log scale. An estimated standard error for $\log r$ is

$$(3.5) \quad \hat{\sigma}(\log r) = \sqrt{\frac{1 - \hat{\pi}_1}{y_1} + \frac{1 - \hat{\pi}_2}{y_2}}.$$

The Wald interval exponentiates endpoints of $\log r \pm z_{\alpha/2} \hat{\sigma}(\log r)$. It tends to be somewhat conservative.

3.1.4 Example: Aspirin and Heart Attacks Revisited

We consider again [Table 2.1](#) from the Harvard study on aspirin use and heart attacks. The proportions having fatal heart attacks were $18/11,034 = 0.00163$ for those taking placebo and $5/11,037 = 0.00045$ for those taking aspirin. The sample relative risk is $0.00163/0.00045 = 3.60$. The 95% confidence interval for the log relative risk, using $\hat{\sigma}(\log r) = 0.505$, is $\log(3.60) \pm 1.96(0.505)$. This translates to $(1.34, 9.70)$ for the relative risk. We infer that the death rate for those taking placebo was between 1.34 and 9.70 times that for those taking aspirin. Substantial public health benefits could result from taking aspirin, but the estimated effect is imprecise despite the very large sample sizes because of the very low rate of heart attack deaths over the study period, regardless of treatment.

The Wald 95% confidence interval for $\pi_1 - \pi_2$ is $0.0012 \pm 1.96(0.00043)$ or $(0.0003, 0.0020)$. The relative risk is more useful than $\pi_1 - \pi_2$ for these data, because the rates of heart attack death were both very low but with ratio quite far from 1.0.

3.1.5 Deriving Standard Errors with the Delta Method

A simple and useful method exists of deriving standard errors. Let T_n denote a statistic that is asymptotically normally distributed about a parameter θ , the subscript n expressing its dependence on sample size. Suppose that an estimator is a function $g(T_n)$ of T_n . Then, under mild conditions, $g(T_n)$ itself has a large-sample normal distribution. The standard error depends on the rate of change of $g(t)$ at $t = \theta$.

Specifically, for large n , suppose that T_n is normally distributed about θ with standard error σ/\sqrt{n} . That is, as $n \rightarrow \infty$, the cdf of $\sqrt{n}(T_n - \theta)$ converges to the cdf of a normal random variable with mean 0 and variance σ^2 . This limiting behavior is an example of *convergence in distribution*, denoted by

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2).$$

Let g be a function that is at least twice differentiable at θ . From the Taylor series expansion for $g(t)$ in a neighborhood of $t = \theta$,

$$\sqrt{n}[g(T_n) - g(\theta)] \approx \sqrt{n}(T_n - \theta)g'(\theta)$$

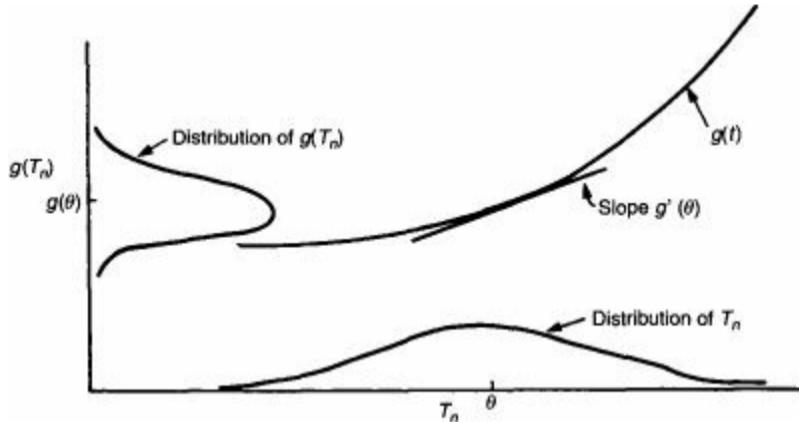
for large n , where $g'(\theta) = \partial g / \partial t$ evaluated at $t = \theta$. Recall if a variate $Y \sim N(0, \sigma^2)$, then $cY \sim N(0, c^2\sigma^2)$. Thus,

$$(3.6) \quad \sqrt{n}[g(T_n) - g(\theta)] \xrightarrow{d} N(0, [g'(\theta)]^2\sigma^2).$$

In other words, $g(T_n)$ is approximately normal around $g(\theta)$ with variance $[g'(\theta)]^2\sigma^2/n$. Section 16.1.2 gives details.

[Figure 3.1](#) portrays this result. Locally around θ , $g(t)$ is approximately linear, with slope $g'(\theta)$. Then $g(T_n)$ is approximately normal, since linear transformations of normal random variables are themselves normal. The dispersion of $g(T_n)$ values about $g(\theta)$ is about $|g'(\theta)|$ times the dispersion of T_n values about θ . For example, if the slope of g at θ is $\frac{1}{2}$, then g maps a region of T_n values into a region of $g(T_n)$ values only about half as wide.

[Figure 3.1](#) Depiction of delta method.



Result (3.6) is called the *delta method*. Since $g'(\theta)$ and $\sigma^2 = \sigma^2(\theta)$ usually depend on the unknown parameter θ , the asymptotic variance is unknown. Wald confidence intervals substitute T_n for θ and use the result that $\sqrt{n}[g(T_n) - g(\theta)]/|g'(T_n)|\sigma(T_n)$ is asymptotically standard normal. Thus,

$$g(T_n) \pm 1.96|g'(T_n)|\sigma(T_n)/\sqrt{n}$$

is a large-sample Wald 95% confidence interval for $g(\theta)$.

3.1.6 Delta Method Applied to the Sample Logit

We illustrate the delta method for a function of the ML estimator $T_n = \hat{\pi} = y/n$ of the binomial parameter π , for y successes in n trials. Recall that $E(\hat{\pi}) = \pi$ and $\text{var}(\hat{\pi}) = \pi(1 - \pi)/n$. Also, $\hat{\pi}$ has a large-sample normal distribution by the central limit theorem. So do many functions of $\hat{\pi}$.

The log odds function of $\hat{\pi}$,

$$g(\hat{\pi}) = \log[\hat{\pi}/(1 - \hat{\pi})],$$

is called the sample *logit*. Evaluated at π , its derivative equals $1/\pi(1 - \pi)$. By the delta method, the asymptotic variance of the sample logit is $\pi(1 - \pi)/n$ (which is the variance of $\hat{\pi}$) multiplied by the square of $[1/\pi(1 - \pi)]$. That is,

$$\sqrt{n} \left(\log \frac{\hat{\pi}}{1 - \hat{\pi}} - \log \frac{\pi}{1 - \pi} \right) \xrightarrow{d} N\left(0, \frac{1}{\pi(1 - \pi)}\right).$$

The asymptotic normality of $\hat{\pi}$ propagates to asymptotic normality of $\log[\hat{\pi}/(1 - \hat{\pi})]$.

The asymptotic variance is the variance of the normal distribution that approximates the true distribution, for large n . It is *not* an approximation for the variance of the true distribution. For $0 < \pi < 1$, the asymptotic variance $[n\pi(1 - \pi)]^{-1}$ of the sample logit is finite. By contrast, the true variance does not exist: Since $\hat{\pi} = 0$ or 1 with positive probability, the logit can equal $-\infty$ or ∞ with positive probability. The probability of an infinite logit converges to zero rapidly as n increases. For large n , the distribution of the sample logit looks essentially normal with mean $\log[\pi/(1 - \pi)]$ and standard deviation $[n\pi(1 - \pi)]^{-1/2}$. Thus, for the logit, the asymptotic variance actually has greater use than the true variance. Incidentally, related to this, the ordinary bootstrap is not helpful for approximating standard errors for many discrete measures, because it mimics the true rather than the more relevant asymptotic standard error.

3.1.7 Delta Method for the Log Odds Ratio

Standard errors for the log odds ratio and the log relative risk result from a multiparameter version of the delta method. Suppose that $\{n_i, i = 1, \dots, c\}$ have a multinomial $(n, \{\pi_i\})$ distribution. The sample proportion $\hat{\pi}_i = n_i/n$ has mean and variance

$$(3.7) E(\hat{\pi}_i) = \pi_i \quad \text{and} \quad \text{var}(\hat{\pi}_i) = \pi_i(1 - \pi_i)/n.$$

In Section 16.1.4 we show that for $i \neq j$, $\hat{\pi}_i$ and $\hat{\pi}_j$ have covariance

$$(3.8) \text{cov}(\hat{\pi}_i, \hat{\pi}_j) = -\pi_i \pi_j / n.$$

The sample proportions $(\hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_{c-1})$ have a large-sample multivariate normal distribution. For functions of them, the delta method implies the following result, proved in Section 16.1.4:

Let $g(\boldsymbol{\pi})$ denote a differentiable function of $\{\pi_i\}$, with sample value $g(\hat{\boldsymbol{\pi}})$ for a multinomial sample. Let

$$\phi_i = \frac{\partial g(\boldsymbol{\pi})}{\partial \pi_i}, \quad i = 1, \dots, c.$$

Then as $n \rightarrow \infty$, the distribution of $\sqrt{n}[g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi})]/\sigma$ converges to standard normal, where

$$(3.9) \sigma^2 = \sum \pi_i \phi_i^2 - \left(\sum \pi_i \phi_i \right)^2.$$

The asymptotic variance depends on $\{\pi_i\}$ and the partial derivatives of the measure with respect to $\{\pi_i\}$. In practice, replacing $\{\pi_i\}$ and $\{\phi_i\}$ in (3.9) by their sample values yields an ML estimate $\hat{\sigma}^2$ of σ^2 . Then $\hat{\sigma}/\sqrt{n}$ is an estimated standard error for $g(\hat{\boldsymbol{\pi}})$. A large-sample Wald confidence interval for $g(\boldsymbol{\pi})$ is

$$g(\hat{\boldsymbol{\pi}}) \pm z_{\alpha/2} \hat{\sigma} / \sqrt{n}.$$

With the substitution of $\hat{\sigma}$ for σ in (3.9), the limiting distribution is still standard normal, but convergence is slower. The equivalence in the large-sample distribution is justified as follows: The sample proportions converge in probability to $\{\pi_i\}$, by the weak law of large numbers. Since $\hat{\sigma}$ is a continuous function of the sample proportions, it converges in probability to σ , and $\sigma / \hat{\sigma}$ converges in probability to 1. Now

$$\sqrt{n} \frac{g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi})}{\hat{\sigma}} = \sqrt{n} \frac{g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi})}{\sigma} \frac{\sigma}{\hat{\sigma}}.$$

The first term on the right-hand side converges in distribution to standard normal, by (3.9), and the second term converges in probability to 1. Thus, their product also has a limiting standard normal distribution.

We now apply the delta method to the log odds ratio, taking $g(\boldsymbol{\pi}) = \log \theta = \log \pi_{11} + \log \pi_{22} - \log \pi_{12} - \log \pi_{21}$. Since

$$\phi_{11} = \partial(\log \theta)/\partial \pi_{11} = 1/\pi_{11}$$

$$\phi_{12} = -1/\pi_{12}, \quad \phi_{21} = -1/\pi_{21}, \quad \phi_{22} = 1/\pi_{22},$$

$\sum_i \sum_j \pi_{ij} \phi_{ij} = 0$ and $\sigma^2 = \sum_i \sum_j \pi_{ij} \phi_{ij}^2 = \sum_i \sum_j (1/\pi_{ij})$. The standard error of $\log \theta$ for a multinomial sample $\{n_{ij}\}$ is

$$\sigma(\log \hat{\theta}) = \sigma/\sqrt{n} = \sqrt{\sum_i \sum_j 1/(n\pi_{ij})}.$$

Since $n_{\hat{\pi}ij} = n_{ij}$, the estimated standard error is (3.1).

3.1.8 Simultaneous Confidence Intervals for Multiple Comparisons

Often, such as in many genetics applications, there are several groups to compare in terms of some parameter. *Multiple comparison* methods apply the confidence level to the simultaneous set of all comparisons, rather than to each individual one.

A simple multipurpose although somewhat conservative way to establish control over a family of inferences is the *Bonferroni method*. For it, with g inferences we use an error probability of $\alpha^* = \alpha/g$ for each one. For instance, to form g confidence intervals with simultaneous coverage probability of at least $1 - \alpha$, we use a standard method but with confidence level $1 - \alpha/g$ for each. This implies an upper bound of α for the probability of at least one error for the entire set of intervals. Exercise 1.36 applied the method to simultaneous comparison of all pairs of multinomial parameters. Goodman (1964a) presented simultaneous confidence intervals for all odds ratios in an $I \times J$ table. Note 3.2 cites an alternative method for comparing multiple binomial parameters. Section 7.5.2 further describes the Bonferroni method, and Section 7.5.3 presents a less conservative approach to multiple comparisons in the context of significance testing.

3.2 TESTING INDEPENDENCE IN TWO-WAY CONTINGENCY TABLES

At first we assume multinomial sampling with joint probabilities $\{\pi_{ij}\}$ in an $I \times J$ contingency table. The null hypothesis of statistical independence is $H_0: \pi_{ij} = \pi_{i+} \pi_{+j}$ for all i and j .

3.2.1 Pearson and Likelihood-Ratio Chi-Squared Tests

In Section 1.5.2 we introduced the Pearson X^2 statistic (1.16) for tests about specified values of multinomial probabilities. A test of H_0 : independence uses X_2 with n_{ij} in place of n_i , and with $\mu_{ij} = n\pi_{i+}\pi_{+j}$ in place of μ_i . Here $\mu_{ij} = E(n_{ij})$ under H_0 . Usually, $\{\pi_{i+}\}$ and $[\pi_{+j}]$ are unknown. Their ML estimates are the sample marginal proportions $\hat{\pi}_{i+} = n_{i+}/n$ and $\hat{\pi}_{+j} = n_{+j}/n$. So, the estimated expected frequencies are $\{\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = n_{i+}n_{+j}/n\}$. Then, the Pearson statistic is

$$(3.10) \quad X^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}.$$

Pearson (1900, 1904, 1922) claimed that replacing $\{\mu_{ij}\}$ by estimates $\{\hat{\mu}_{ij}\}$ would not affect the large-sample distribution of X^2 . Since the contingency table has IJ categories, he argued that X^2 is asymptotically chi-squared with $df = IJ - 1$. On the contrary, since $\{\hat{\mu}_{ij}\}$ require estimating $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$, by Section 1.5.6,

$$df = (IJ - 1) - (I - 1) - (J - 1) = (I - 1)(J - 1).$$

The dimensions of $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ reflect the constraints $\sum_i \pi_{i+} = \sum_j \pi_{+j} = 1$. R. A. Fisher (1922) corrected Pearson's error. Fisher's article introduced the notion of *degrees of freedom*. (Pearson had introduced an indexed family of chi-squared distributions but had not dealt explicitly with "degrees of freedom.")

The score test produces the X^2 statistic. The likelihood-ratio test produces a different statistic. For multinomial sampling, the kernel of the likelihood is

$$\prod_i \prod_j \pi_{ij}^{n_{ij}}, \quad \text{where all } \pi_{ij} \geq 0 \quad \text{and} \quad \sum_i \sum_j \pi_{ij} = 1.$$

Under H_0 : independence, $\hat{\pi}_{ij} = \hat{\pi}_{i+}\hat{\pi}_{+j} = n_{i+}n_{+j}/n^2$. In the general case, $\hat{\pi}_{ij} = n_{ij}/n$. The ratio of the likelihoods equals

$$\Lambda = \frac{\prod_i \prod_j (n_{i+}n_{+j})^{n_{ij}}}{n^n \prod_i \prod_j n_{ij}^{n_{ij}}}.$$

The likelihood-ratio chi-squared statistic is $-2 \log \Lambda$. Denoted by G^2 , it equals

$$(3.11) \quad G^2 = -2 \log \Lambda = 2 \sum_i \sum_j n_{ij} \log(n_{ij}/\hat{\mu}_{ij}),$$

The larger the values of G^2 and X^2 , the more evidence exists against independence. For either statistic, the P -value is the right-tail probability above the observed value.

In the general case, the parameter space consists of $\{\pi_{ij}\}$ subject to the linear restriction $\sum_i \sum_j \pi_{ij} = 1$, so the dimension is $IJ - 1$. Under H_0 , $\{\pi_{ij}\}$ are determined by $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$, so the dimension is $(I - 1) + (J - 1)$. The difference in these dimensions equals $(I - 1)(J - 1)$. For large samples, G^2 has a chi-squared null distribution with $df = (I - 1)(J - 1)$. So G^2 and X^2 have the same limiting null chi-squared distribution. In fact, they are then asymptotically equivalent; $X^2 - G^2$ converges in probability to zero (Section 16.3.4).

When there are independent multinomial samples in the I rows, the row marginal counts are fixed. Independence then corresponds to homogeneity of each outcome probability among the rows. Roy and Mitra (1956) showed that the limiting chi-squared results for a single multinomial sample also hold then (and for comparable statistics in three-way tables), as well as when we condition further on the column marginal totals. As we'll discuss in Section 3.5, conditional on row and column marginal totals, a hypergeometric distribution applies to the cell counts. In this case, $\{\hat{\mu}_{ij}\}$ in tests of independence are *exact* (rather than *estimated*) expected values. For 2×2 tables, for example,

$$E(n_{11}) = \frac{n_{1+}n_{+1}}{n} \quad \text{and} \quad \text{var}(n_{11}) = \frac{n_{1+}n_{+1}n_{2+}n_{+2}}{n^2(n-1)}.$$

For $I \times J$ tables, Haldane (1940) derived $E(X^2) = (I - 1)(J - 1)n/(n - 1)$. See Note 3.3 for other moments.

3.2.2 Example: Education and Belief in God

[Table 3.2](#) uses General Social Survey data to cross-classify opinion about whether God exists by highest education degree attained. The table also contains the estimated expected frequencies for H_0 : independence. For instance, $\hat{\mu}_{11} = n_{1+} n_{+1}/n = (335 \times 60)/2000 = 10.0$. The chi-squared statistics are $X^2 = 76.1$ and $G^2 = 73.2$, with $df = (3 - 1)(6 - 1) = 10$. The P -values are < 0.0001 . These statistics provide extremely strong evidence of an association.

Table 3.2 Attained Education (Highest Degree) and Belief in God

Highest Degree	Belief in God							Total
	Don't Believe	No Way to Find Out	Some Higher Power	Believe Sometimes	Believe but Doubts	Know God Exists		
Less than high school	9 (10.0) ^a (-0.4) ^b	8 (15.9) (-2.2)	27 (34.2) (-1.4)	8 (12.7) (-1.5)	47 (55.3) (-1.3)	236 (206.9) (3.6)	335	
High school or junior college	23 (32.5) (-2.5)	39 (51.5) (-2.6)	88 (110.6) (-3.3)	49 (41.2) (1.8)	179 (178.9) (0.0)	706 (669.4) (3.4)	1084	
Bachelor or graduate	28 (17.4) (3.1)	48 (27.6) (4.7)	89 (59.3) (4.8)	19 (22.1) (-0.8)	104 (95.9) (1.1)	293 (358.8) (-6.7)	581	
Total	60	95	204	76	330	1235	2000	

^aEstimated expected frequencies for testing independence.

^bStandardized residuals.

Source: 2008 General Social Survey, National Opinion Research Center.

3.2.3 Adequacy of Chi-Squared Approximations

The convergence of the actual sampling distribution of X^2 or G^2 to the chi-squared distribution applies as n grows, and hence $\{\mu_{ij} = n\pi_{ij}\}$ grow, for a fixed number of cells. As the cell means grow, the multinomial distribution for $\{n_{ij}\}$ is better approximated by a multivariate normal, and X^2 and G^2 have more nearly chi-squared distributions. The adequacy of the approximation depends on both n and the number of cells. The size of n/IJ that produces adequate approximations for X^2 tends to decrease as IJ increases (Koehler and Larntz 1980).

Contingency tables having small cell counts are said to be *sparse*. In analyzing the chi-squared approximation for X^2 in sparse tables, Cochran (1954) suggested that when $df > 1$, a minimum expected value $\mu_{ij} \geq 1$ is permissible as long as no more than about 20% of $\mu_{ij} < 5$. Research has shown that X^2 performs adequately with smaller n and more sparse tables than G^2 (see Note 3.3). The distribution of G^2 is usually poorly approximated by chi-squared when $n/IJ < 5$. Depending on the sparseness, P -values based on referring G^2 to a chi-squared distribution can be too large or too small. When most μ_{ij} are smaller than 0.50, treating G^2 as chi-squared gives a highly conservative test; when H_0 is true, reported P -values tend to be much larger than true ones. When most μ_{ij} are between 0.5 and 4, by contrast, the reported P -value tends to be too small.

A caveat is that chi-squared approximations tend to be poor for tables containing both very small and moderately large μ_{ij} (Haberman 1988). It is difficult to give a guideline that covers all cases. Small-sample methods to be presented in Section 3.5 are available whenever it is doubtful whether n is sufficiently large.

3.2.4 Chi-Squared and Comparing Proportions in 2×2 Tables

Often, a 2×2 table summarizes results for two independent binomial variates y_1 and y_2 with n_1 and n_2 trials. Independence is equivalent to the homogeneity condition, $\pi_1 = \pi_2$. Under $H_0: \pi_1 = \pi_2$, the estimated common value of $\pi_1 = \pi_2$ is $\hat{\pi} = (y_1 + y_2)/(n_1 + n_2)$. The z score test statistic

$$(3.12) \quad z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1 - \hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

has denominator that is the standard error of $\hat{\pi}_1 - \hat{\pi}_2$ estimated under H_0 . This statistic has an asymptotic standard normal null distribution.

This statistic relates to the Pearson statistic for testing independence in the 2×2 table by $z^2 = X^2$. Recall that if a statistic z has an approximate standard normal distribution, then z^2 has an approximate chi-squared distribution with $df = 1$, which is $(I - 1)(J - 1)$ applied with $I = J = 2$.

A simple formula for X^2 for 2×2 tables is

$$X^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}}.$$

For example, for the 2×2 table having entries (3, 0 / 0, 3), by row, used for an example in Section 3.5.6 on small-sample inference,

$$X^2 = [6(3 \times 3 - 0 \times 0)^2]/(3 \times 3 \times 3 \times 3) = 6.0.$$

Section 5.3.5 shows a generalized formula for comparing I proportions in $I \times 2$ tables. Mirkin (2001) showed alternative X^2 formulas for $I \times J$ tables.

3.2.5 Score Confidence Intervals Comparing Proportions

The Wald confidence intervals for the difference of proportions, odds ratio, and relative risk presented in Section 3.1 are simple but have disadvantages: They are dependent on the scale of measurement [e.g., a Wald interval is not the same for θ as when found for $\log(\theta)$ and then exponentiated], they fail when an estimate falls at the boundary of the parameter space [e.g., a cell count of 0 causing $\log(\theta) = \pm\infty$ and $\hat{\sigma}(\log \theta) = \infty$], and they can have actual probability of covering the parameter quite far from the nominal level unless n is quite large. Alternative intervals that result from inverting score tests or likelihood-ratio tests do not have these disadvantages. These tests use extensions of the X^2 or G^2 statistics that apply to nonnull values of the parameters. Although computationally more complex than the Wald method, this should not be an impediment to their use in this modern era of computing, as the principle behind them is straightforward.

We illustrate the score method for forming an interval for the difference of proportions. Consider testing $H_0: \pi_1 - \pi_2 = \Delta_0$, where Δ_0 need not be 0. Let $\hat{\pi}_1(\Delta_0)$ and $\hat{\pi}_2(\Delta_0)$ denote the ML estimates of π_1 and π_2 subject to the constraint $\pi_1 - \pi_2 = \Delta_0$. That is, $\hat{\pi}_1(\Delta_0)$ and $\hat{\pi}_2(\Delta_0)$ are the values of π_1 and π_2 satisfying $\pi_1 - \pi_2 = \Delta_0$ that maximize the product of the two binomial probability mass functions. The score test statistic is

$$z(\Delta_0) = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - \Delta_0}{\sqrt{\frac{\hat{\pi}_1(\Delta_0)[1 - \hat{\pi}_1(\Delta_0)]}{n_1} + \frac{\hat{\pi}_2(\Delta_0)[1 - \hat{\pi}_2(\Delta_0)]}{n_2}}}.$$

The score confidence interval is the set of Δ_0 such that $|z(\Delta_0)| < z_{\alpha/2}$ (Mee 1984). For given Δ_0 , each $\hat{\pi}_i(\Delta_0)$ and hence $z(\Delta_0)$ can be found explicitly, but finding the endpoints of the interval requires iteration (Nurminen 1986).

For $\Delta_0 = 0$, the test statistic $z(\Delta_0)$ simplifies to the pooled statistic (3.12) for comparing two proportions. Then, $[z(\Delta_0)]^2$ is the Pearson X^2 statistic. For $\Delta_0 \neq 0$ this square is a nonnull type of Pearson statistic. Unlike the Wald interval, the score interval is coherent with the result of the Pearson chi-squared test of independence; for instance, the P -value for that test falling below 0.05 is equivalent to the 95% score confidence interval for $\pi_1 - \pi_2$ not containing 0. For Table 2.1 on aspirin use and heart attacks, the 95% score interval for $\pi_1 - \pi_2$ is (0.0004, 0.0022).

Score-test-based confidence intervals have also been proposed for the odds ratio (Cornfield 1956) and for the relative risk (Koopman 1984). We illustrate for the odds ratio for a multinomial sample over the cells of the 2×2 table. Recall that the joint distribution $\{\pi_{ij}\}$ can equivalently be expressed in terms of $\{\theta, \pi_{1+}, \pi_{+1}\}$ (Section 2.4.1). For a given nonnull odds ratio value θ_0 , let $\{\hat{\mu}_{ij}(\theta_0)\}$ be the unique expected frequency estimates that have the same row and column margins as $\{n_{ij}\}$ and satisfy

$$\frac{\hat{\mu}_{11}(\theta_0)\hat{\mu}_{22}(\theta_0)}{\hat{\mu}_{12}(\theta_0)\hat{\mu}_{21}(\theta_0)} = \theta_0.$$

The set of θ_0 satisfying

$$X^2(\theta_0) = \sum (n_{ij} - \hat{\mu}_{ij}(\theta_0))^2 / \hat{\mu}_{ij}(\theta_0) < \chi^2_1(\alpha)$$

form a $100(1 - \alpha)\%$ score-test-based confidence interval. This interval is also coherent with the result of the Pearson chi-squared test, for $H_0: \theta = 1$. This 95% score interval for the odds ratio for Table 3.1 on seat-belt use and traffic accidents is (6.76, 17.35).

3.2.6 Profile Likelihood Confidence Intervals

Likewise, we can construct confidence intervals by inverting likelihood-ratio tests for nonnull parameter values. We illustrate with the odds ratio. For $\{\hat{\mu}_{ij}(\theta_0)\}$ as just defined, the set of θ_0 satisfying

$$G^2(\theta_0) = 2 \sum_i \sum_j n_{ij} \log[n_{ij}/\hat{\mu}_{ij}(\theta_0)] < \chi_1^2(\alpha)$$

form a $100(1 - \alpha)\%$ likelihood-ratio test-based confidence interval. The 95% interval for the odds ratio for [Table 3.1](#) on seat-belt use and traffic accidents is (6.82, 17.70).

More generally, in later chapters we'll often construct a confidence interval for a model parameter β , regarding the other parameters in the model as *nuisance parameters*. Denote those nuisance parameters, such as the marginal probabilities in a 2×2 table when we are estimating an odds ratio, by ψ . In inverting a likelihood-ratio test of $H_0: \beta = \beta_0$ to check whether β_0 belongs in the confidence interval, the ML estimate $\hat{\psi}(\beta_0)$ that maximizes the likelihood under the null varies as β_0 does. The *profile log-likelihood function* is $L(\beta_0, \hat{\psi}(\beta_0))$, viewed as a function of β_0 . For each β_0 this function gives the maximum of the ordinary log likelihood subject to the constraint $\beta = \beta_0$. Evaluated at $\beta_0 = \hat{\beta}$, this is the maximized log likelihood $L(\hat{\beta}, \hat{\psi})$, which occurs at the unrestricted ML estimates. The *profile likelihood confidence interval* for β is the set of β_0 for which

$$-2[L(\beta_0, \hat{\psi}(\beta_0)) - L(\hat{\beta}, \hat{\psi})] - < \chi_1^2(\alpha).$$

The interval contains all β_0 not rejected in likelihood-ratio tests of nominal size α .

Score intervals currently are available only in specialized software, such as R functions given in this book's computing appendix.² The profile likelihood approach is more generally available, for example the *confint()* function in R, the *LRCI* option in PROC GENMOD and the *PLCL* option in PROC LOGISTIC in SAS, and the *pllf* command in Stata.

3.3 FOLLOWING-UP CHI-SQUARED TESTS

Like any significance test, chi-squared tests of independence have limited usefulness. A small P -value indicates strong evidence of association but provides little information about the nature or strength of the association. Statisticians have long warned about dangers of relying solely on results of chi-squared tests rather than studying the nature of the association (e.g., Berkson 1938, Cochran 1954). In this section we discuss ways to follow up the tests to learn more about the association.

3.3.1 Pearson Residuals and Standardized Residuals

A cell-by-cell comparison of observed and estimated expected frequencies helps show the nature of the dependence. Under H_0 , larger differences $(n_{ij} - \hat{\mu}_{ij})$ tend to occur in cells with larger μ_{ij} . Thus, this raw difference is insufficient. The *Pearson residual*, defined for a cell by

$$(3.13) \quad e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}},$$

attempts to adjust for this. The name “Pearson” results from $\{e_{ij}\}$ relating to the Pearson statistic by $X^2 = \sum_i \sum_j e_{ij}^2$.

Under H_0 , $\{e_{ij}\}$ are asymptotically normal with mean 0. However, their asymptotic variances are less than 1.0, averaging $[(I-1)(J-1)]/IJ$. A *standardized residual* that is asymptotically standard normal results from dividing $(n_{ij} - \hat{\mu}_{ij})$ by its standard error (Haberman 1973a, Sec. 16.3.2). For H_0 : independence, this is

$$(3.14) \quad r_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-p_{i+})(1-p_{+j})}}.$$

In 2×2 tables, $df = 1$ and $r_{11} = -r_{12} = -r_{21} = r_{22}$ and any $r_{ij}^2 = X^2$. By contrast, all four Pearson residuals can take different values, which is unappealing.

A standardized residual that exceeds about 2 or 3 in absolute value indicates lack of fit of H_0 in that cell. Larger values are more relevant when df is larger, as it becomes more likely that at least one such residual is large simply by chance.

3.3.2 Example: Education and Belief in God Revisited

[Table 3.2](#) also shows standardized residuals for testing independence. For instance, $n_{36} = 293$ and $\mu_{36} = 358.8$. The relevant marginal proportions equal $p_{3+} = 581/2000 = 0.2905$ and $p_{+6} = 1235/2000 = 0.6175$. The standardized residual [\(3.14\)](#) for this cell equals

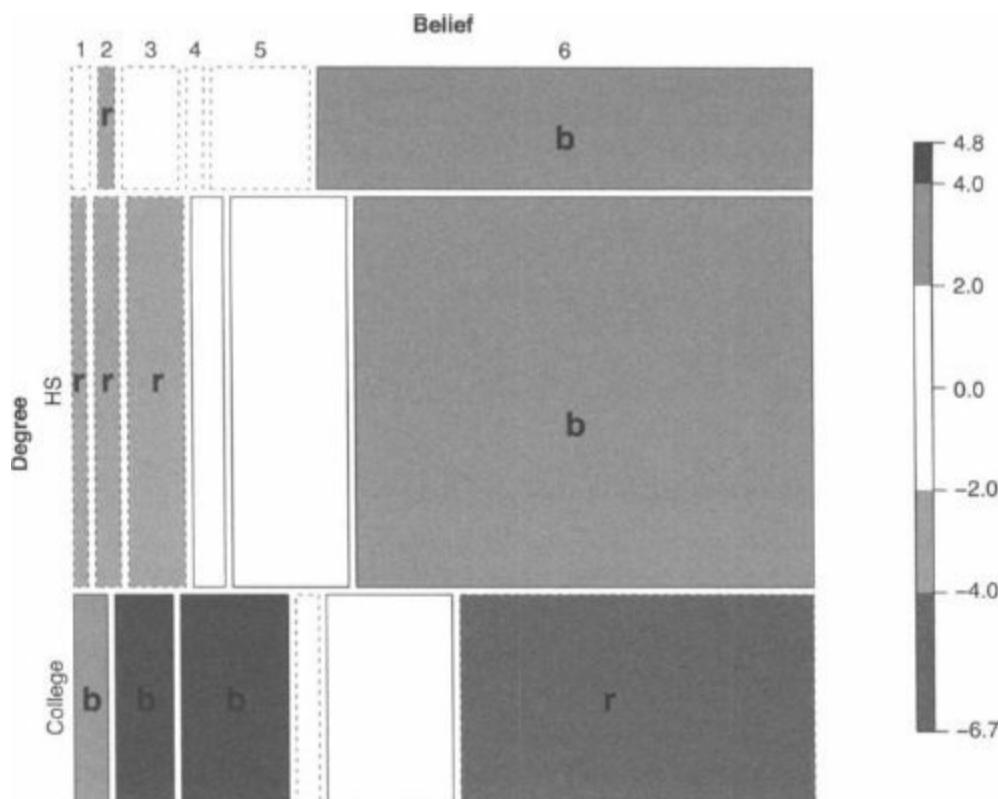
$$r_{36} = (293 - 358.8)/\sqrt{358.8(1 - 0.2905)(1 - 0.6175)} = -6.7.$$

We can infer that, in the population in 2008, fewer people at the highest level of education would have responded “know God exists” than if the variables were truly independent.

For the “know God exists” category, [Table 3.2](#) shows large positive residuals for subjects with a junior college education or less. We can infer that more subjects at these education levels had this opinion than if H_0 : independence were true. Other large positive residuals occur in the first three categories of belief in God for those with at least a bachelor degree, suggesting those cells are also more common than we’d expect under independence.

[Figure 3.2](#) is a *mosaic plot* for [Table 3.2](#). Mosaic plots portray the counts by tiles (rectangles) whose size is proportional to the cell count. Under independence, the vertical lines would match up at the same spot in each row. Color and depth of shading of the tiles can represent the sign and magnitude of standardized residuals (Friendly 1994). The scale on the right of the figure shows the magnitude of the standardized residuals.

[Figure 3.2](#) Mosaic plot for data in [Table 3.2](#). [Figure 3.2](#), when produced with a mosaic() function in R, has blue tiles (labeled b here) for positive residuals and red (labeled r here) for negative, with dark color when the standardized value exceeds 4.



3.3.3 Partitioning Chi-Squared

Another supplement to a chi-squared test uses the reproductive property of chi-squared (Section 1.2.6) to partition the test statistic so that the components represent certain aspects of the effects. A partitioning may show that an association reflects primarily differences between certain categories or groupings of categories.

We begin with a partitioning for the test of independence in $2 \times J$ tables. We partition G^2 , which has $\text{df} = (J - 1)$, into $J - 1$ components. The j th component is G^2 for a 2×2 table where the first column combines columns 1 through j of the full table and the second column is column $j + 1$. That is, G^2 for testing independence in a $2 \times J$ table equals a statistic that compares the first two columns, plus a statistic that combines the first two columns and compares them to the third column, and so on, up to a statistic that combines the first $J - 1$ columns and compares them to the last column.³ Each component statistic has $\text{df} = 1$.

It might seem more natural to compute G^2 for the $(J - 1)$ separate 2×2 tables that pair each column with a particular one, say, the last. Such an analysis can be informative, but these component statistics are not independent and do not sum to G^2 for the full table. This is beyond our scope at this stage but relates to the contrasts of log probabilities that form the log odds ratios for the two tables not being orthogonal.

For an $I \times I$ table, independent chi-squared components result from comparing columns 1 and 2 and then combining them and comparing them to column 3, and so on. Each of the $J - 1$ statistics has $\text{df} = I - 1$. More refined partitions contain $(I - 1)(J - 1)$ statistics, each having $\text{df} = 1$. One such partition (Lancaster 1949a) applies to the $(I - 1)(J - 1)$ separate 2×2 tables

$$(3.15) \quad \begin{array}{c|c} \sum_{a < i} \sum_{b < j} n_{ab} & \sum_{a < i} n_{aj} \\ \hline \sum_{b < j} n_{ib} & n_{ij} \end{array}$$

for $i = 2, \dots, I$ and $j = 2, \dots, J$. For others, see Gilula and Haberman (2005) and Goodman (1969, 1971b).

3.3.4 Example: Origin of Schizophrenia

[Table 3.3](#) classifies a sample of psychiatrists by their school of psychiatric thought and by their opinion on the origin of schizophrenia. Here $G^2 = 23.04$ with $df = 4$. To understand this association better, we partition G^2 into four independent components.

Table 3.3 Most Influential School of Psychiatric Thought and Ascribed Origin of Schizophrenia

School of Psychiatric Thought	Origin of Schizophrenia		
	Biogenic	Environmental	Combination
Eclectic	90	12	78
Medical	13	1	6
Psychoanalytic	19	13	50

Source: Reprinted with permission, based on data from B. J. Gallagher III, B. J. Jones, and L. P. Barakat, *J. Clin. Psychol.* 43: 438–443, 1987.

The partitioning (3.15) applies to the subtables shown in [Table 3.4](#). The first subtable compares the eclectic and medical schools of psychiatric thought on whether the origin of schizophrenia is biogenic or environmental given that the classification was in one of these two categories. For this subtable, $G^2 = 0.29$, with $df = 1$. The second subtable compares these two schools on the proportion of times the origin was ascribed to be a combination, rather than biogenic or environmental. This subtable has $G^2 = 1.36$, with $df = 1$. The sum of these two components equals G^2 for testing independence with the first two rows of [Table 3.3](#). There is little evidence of a difference between the eclectic and medical schools of thought on the ascribed origin of schizophrenia.

Table 3.4 Subtables Used in Partitioning Chi-Squared for [Table 3.3](#)^a

		Bio + Env		Bio Env		Bio + Env	
Bio	Env	Env	Com	Bio	Env	Env	Com
Ecl	90	12	Ecl	102	78	Ecl + Med	103
Med	13	1	Med	14	6	Psy	19

		Bio + Env		Bio Env		Bio + Env	
Bio	Env	Env	Com	Bio	Env	Env	Com
Ecl	90	12	Ecl	102	78	Ecl + Med	103
Med	13	1	Med	14	6	Psy	19

^aBio, biogenic; Com, combination; Ecl, eclectic; Env, environmental; Psy, psychoanalytic.

Next, we combine the eclectic and medical schools and compare them to the psychoanalytic school. The third subtable in [Table 3.4](#) compares them for the (biogenic, environmental) classification, giving $G^2 = 12.95$ with $df = 1$. The fourth subtable compares them for the (biogenic or environmental, combination) split, giving $G^2 = 8.43$ with $df = 1$.

The psychoanalytic school seems more likely than the other schools to ascribe the origins of schizophrenia as being a combination. Of those who chose either the biogenic or environmental origin, members of the psychoanalytic school were somewhat more likely than the other schools to choose the environmental origin. The sum of these four G^2 components equals the value of 23.04 for testing independence in the full 3×3 table.

3.3.5 Rules for Partitioning

Goodman (1968, 1969, 1971b) and Lancaster (1949a, 1969) gave rules for determining independent components of chi-squared. For forming subtables, among the necessary conditions are the following:

1. The df for the subtables must sum to the df for the full table.
2. Each cell count in the full table must be a cell count in one and only one subtable.
3. Each marginal total of the full table must be a marginal total for one and only one subtable.

For a certain partitioning, when the subtable df values sum properly but the G^2 values do not, the components are not independent.

For the G^2 statistic, exact partitionings occur. The Pearson X^2 need not equal the sum of the X^2 values for the subtables. It is valid to use the X^2 statistics for the separate subtables; they simply do not provide an exact algebraic partitioning of X^2 for the full table. When the null hypotheses all hold, X^2 does have an asymptotic equivalence with G^2 , however. In addition, when the table has small counts and we rely on large-sample distributions, it is safer to use X^2 than G^2 to analyze the subtables.

3.3.6 Summarizing the Association

Residual analyses and partitioning of chi-squared are both inferential methods. They provide information about whether there is an association and its nature, but in an inferential manner. For example, as n increases and there truly is an association, standardized residuals tend to be larger in magnitude, but they do not describe the strength of association.

To describe the strength of association, we can use measures introduced in the previous chapter, such as the odds ratio, by applying them to either subtables or collapsings of the table. We illustrate with [Table 3.2](#) on education and belief in God. The 2×2 table constructed by combining the first two rows and combining the first five columns has a sample odds ratio of $(477 \times 293)/(942 \times 288) = 0.52$. For those with at least a bachelor's degree, the estimated odds of responding "know God exists" were 0.52 times the estimated odds for those with less than a bachelor's degree. Likewise, we can use measures such as differences and ratios of proportions. For example, the sample proportion responding "know God exists" was 0.704 for those with less than a high school education and 0.504 for those with a bachelor's degree or higher, for a difference of 0.20 and a ratio of 1.40. We can also construct confidence intervals for such parameters, as discussed in Sections 3.1, 3.2.5, and 3.2.6.

A useful summary of the degree to which cells depart from independence compares cell counts with the independence fit by the estimates $\{a_{ij} = n_{ij}/\mu_{ij} = p_{ij}/(p_{i+}p_{+j})\}$ of the association factors (Section 2.4.2). For those with the highest degree who responded "know God exists," this is $a_{36} = 293/[(581)(1235)/2000] = 0.82$; that is, the observed count was 82% of what independence predicts.

3.3.7 Limitations of Chi-Squared Tests

Chi-squared tests of independence merely indicate the degree of evidence of association. They are rarely adequate for answering all questions about a data set. Rather than relying solely on results of these tests, investigate the nature of the association: Look at the standardized residuals, decompose chi-squared into components, and estimate parameters that describe the strength of association.

The chi-squared tests also have limitations in the types of data to which they apply. For instance, they require large samples. Also, $\{\mu_{ij} = n_{i+} n_{+j}/n\}$ used in X^2 and G^2 depend on the marginal totals but not on the order of listing the rows and columns. Thus, X^2 and G^2 do not change value with arbitrary reorderings of rows or of columns. This implies that they treat both classifications as nominal. When at least one variable is ordinal, test statistics that utilize the ordinality are usually more appropriate. We present such tests in Section 3.4.

3.3.8 Why Consider Independence If It's Unlikely to Be True?

Any idealized structure such as independence is unlikely to hold in many situations. With large samples such as in [Table 3.2](#), it is not surprising to obtain a small P -value. Given this and the limitations just mentioned, why even bother to consider independence as a possible representation for a joint distribution?

One reason refers to the benefits of *parsimony*, using fewer parameters to describe the data. The estimates $\{\hat{\pi}_{ij} = n_{i+}n_{+j}/n^2\}$ of the cell probabilities are based on estimating the $(I - 1) + (J - 1)$ marginal probability parameters $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$. By contrast, the sample proportions $\{p_{ij} = n_{ij}/n\}$ are based on estimating the $IJ - 1$ cell probability parameters $\{\pi_{ij}\}$. When the independence hypothesis approximates the true probabilities well, unless n is very large the independence-based ML estimates tend to be better than the sample proportions. The independence estimates smooth the sample counts, somewhat damping the random sampling fluctuations. This is the same reason that we use models to smooth data in the rest of the text.

The mean squared error (MSE) formula

$$\text{MSE} = \text{variance} + (\text{bias})^2$$

explains why the independence estimators can have smaller MSE. Although they may be biased, they have smaller variance because they are based on estimating fewer parameters. Hence, MSE can be smaller unless n is so large that the bias term dominates the variance.

We illustrate using [Table 3.5](#), which has $\pi_{ij} = \pi_{i+}\pi_{+j}[1 + \delta(i - 2)(j - 2)]$ for $\pi_{i+} = \pi_{+j} = \frac{1}{3}$. Here $-1 < \delta < 1$, with $\delta = 0$ equivalent to independence. When δ is close to zero, independence approximates the relationship well. The total MSE values of the two estimators are

Table 3.5 Cell Probabilities for MSE Comparison of Estimators

$(1+\delta)/9$	$1/9$	$(1-\delta)/9$
$1/9$	$1/9$	$1/9$
$(1-\delta)/9$	$1/9$	$(1+\delta)/9$

$$\begin{aligned} \text{MSE}(\{p_{ij}\}) &= \sum_i \sum_j E(p_{ij} - \pi_{ij})^2 = \sum_i \sum_j \text{var}(p_{ij}) \\ &= \sum_i \sum_j \pi_{ij}(1 - \pi_{ij})/n = \frac{1}{n} \left(1 - \sum_i \sum_j \pi_{ij}^2 \right), \end{aligned}$$

$$\text{MSE}(\{\hat{\pi}_{ij}\}) = \sum_i \sum_j E(\hat{\pi}_{ij} - \pi_{ij})^2.$$

For [Table 3.5](#),

$$\text{MSE}(\{p_{ij}\}) = \frac{1}{n} \left(\frac{8}{9} - \frac{4\delta^2}{81} \right)$$

and rather tedious calculations yield

$$\text{MSE}(\{\hat{\pi}_{ij}\}) = \frac{1}{n} \left(\frac{4}{9} + \frac{4}{9n} \right) + \frac{4\delta^2}{81} \left(1 - \frac{2}{n} + \frac{2}{n^2} - \frac{2}{n^3} \right).$$

[Table 3.6](#) lists the total MSE values for various δ and n . When $\delta = 0$, $\text{MSE}(\{p_{ij}\}) = 8/9n$, whereas $\text{MSE}(\{\hat{\pi}_{ij}\}) \approx 4/9n$ for large n . The independence estimator is then much better. When the table is close to independence ($\delta \approx 0$) and n is not large, MSE is only about half as large for the independence estimator. When $\delta \neq 0$, the inconsistency of $\{\hat{\pi}_{ij}\}$ is reflected by $\text{MSE}(\{\hat{\pi}_{ij}\}) \rightarrow 4\delta^2/81$ [whereas $\text{MSE}(\{p_{ij}\}) \rightarrow 0$] as $n \rightarrow \infty$. When the table is close to independence, however, the independence estimator has a lower total MSE even for moderately large n (e.g., for $n = 500$ when $\delta = 0.1$).

Table 3.6 Comparison of Total MSE(x10,000) for Sample Proportion (p_{ij}) and Independence ($\hat{\pi}_{ij}$) Estimators of the Cell Probabilities in [Table 3.5](#)

n	$\delta = 0$		$\delta = 0.1$		$\delta = 0.2$		$\delta = 0.6$		$\delta = 1.0$	
	p	$\hat{\pi}$	p	$\hat{\pi}$	p	$\hat{\pi}$	p	$\hat{\pi}$	p	$\hat{\pi}$
10	889	489	888	493	887	505	871	634	840	893
50	178	91	178	95	177	110	174	261	168	565
100	89	45	89	50	89	65	87	220	84	529
500	18	9	18	14	18	28	17	186	17	500
∞	0	0	0	5	0	20	0	178	0	494

3.4 TWO-WAY TABLES WITH ORDERED CLASSIFICATIONS

The X^2 and G^2 chi-squared tests ignore some information when used to test independence between ordinal classifications. When rows and/or columns are ordered, other tests that take the ordering into account are usually more powerful.

3.4.1 Linear Trend Alternative to Independence

When the row variable X and the column variable Y are ordinal, a positive or negative trend in the association is common. One approach to inference, described later in this section, uses an ordinal measure of monotone trend. An alternative analysis assigns scores to categories and summarizes the *linear trend* component of the association.

A test statistic that is sensitive to positive or negative linear trends utilizes correlation information. Let $u_1 \leq u_2 \leq \dots \leq u_I$ denote scores for the rows, and let $v_1 \leq v_2 \leq \dots \leq v_J$ denote column scores. The scores have the same ordering as the categories. They assign distances between categories and actually treat the measurement scale as interval, with greater distances between categories that are farther apart.

The sum $\sum_i \sum_j u_i v_j p_{ij}$ weights cross-products of scores by their relative frequency, $p_{ij} = n_{ij}/n$. It relates to the covariation of X and Y . For the scores chosen, the correlation r between X and Y equals the standardization of this sum to the -1 to $+1$ scale. (In fact, r equals this sum when both sets of scores are linearly transformed for the n subjects to have a mean of 0 and standard deviation of 1.) The larger $|r|$ is in absolute value, the farther the data fall from independence in this linear dimension.

A statistic for testing independence against the two-sided alternative of nonzero true correlation is

$$(3.16) M^2 = (n - 1)r^2.$$

This statistic increases as $|r|$ or n does. For large samples, it is approximately chi-squared with $df = 1$ (Mantel 1963, Yates 1948). Large values contradict independence, so as with X^2 and G^2 , the P -value is the right-tail probability above the value observed. A small P -value does not imply that the association is linear, but merely that the linear component of the association is significant. The test treats the variables symmetrically.

3.4.2 Example: Is Happiness Associated with Political Ideology?

[Table 3.7](#) cross-classifies degree of happiness by political ideology for all subjects aged over 65 in the 2008 GSS. The Pearson chi-squared statistics for testing independence is $X^2 = 7.07$ with $df = 4$ (P -value = 0.13). This statistic shows little evidence of association, but it ignores the ordering of rows and columns. With scores (1, 2, 3) for each variable, the correlation is $r = 0.135$. The linear trend test statistic $M^2 = (321 - 1)(0.135)^2 = 5.85$ with $df = 1$. This shows strong evidence of association ($P = 0.016$).

Table 3.7 Happiness and Political Ideology

Political Ideology	Happiness		
	Not too Happy	Pretty Happy	Very Happy
Liberal	13	29	15
Moderate	23	59	47
Conservative	14	67	54

Source: 2008 General Social Survey.

The nontrivial evidence of association may be surprising, since X^2 has such an unimpressive value. When a positive or negative trend exists, analyses designed to detect that trend have greater power and tend to provide smaller P -values than analyses that ignore it.

3.4.3 Monotone Trend Alternatives to Independence

Ordinal variables do not have a specified metric. The method of detecting a linear trend alternative to independence requires assigning scores to X and Y , treating them as interval variables. Alternatively, we can add more structure and perform inference about a correlation for an assumed underlying continuous distribution, as the polychoric correlation does with the normal distribution (Section 2.4.8). In the opposite direction, a strict ordinal analysis with the weaker alternative of monotonicity uses an ordinal measure of association, such as gamma (Section 2.4.5). Inference is available with each of these approaches.

For example, with large random samples, sample gamma has approximately a normal sampling distribution. The standard error follows from the delta method (Goodman and Kruskal 1963). Gamma is the basis of an ordinal test of independence using test statistic $z = \hat{\gamma}/SE$. A confidence interval describes the strength of positive or negative monotone association. It is also possible to use as test statistic the ratio of $(C - D)/SE_0$ for a null standard error obtained under the condition of independence (Agresti 2010, Sec. 7.3.3).

For [Table 3.7](#) on happiness and political ideology, $\hat{\gamma} = 0.185$. The sample has a weak tendency for happiness to increase as political conservatism increases. Software⁴ reports a standard error of 0.078 for gamma. There is considerable evidence that the population value $\gamma > 0$, since $z = 0.185/0.078 = 2.37$ ($P = 0.018$ for the two-sided alternative). An approximate 95% confidence interval for γ is $0.185 \pm 1.96(0.078)$, or $(0.032, 0.338)$. The true association seems to be relatively weak and could be very weak.

3.4.4 Extra Power with Ordinal Tests

For testing independence, X^2 and G^2 refer to the most general alternative, whereby cell probabilities exhibit *any* type of statistical dependence. Their df value of $(I - 1)(J - 1)$ reflects an alternative hypothesis that has $(I - 1)(J - 1)$ more parameters than the null hypothesis—the nonredundant odds ratios that describe the association [such as [\(2.10\)](#)]. These statistics are designed to detect *any* pattern for these parameters. In achieving this generality, they sacrifice sensitivity for detecting particular patterns.

By contrast, the analyses for ordinal row and column variables describe association using a single parameter. For instance, M^2 uses the correlation. When a chi-squared test statistic refers to a single parameter [such as M^2 or $(\hat{\gamma}/SE)^2$ does], it has $df = 1$. When the association truly has a positive or negative trend, an ordinal test has a power advantage over the tests using X^2 or G^2 . Since df equals the mean of the chi-squared distribution, a relatively large M^2 value with $df = 1$ falls farther out in its right-hand tail than a comparable value of X^2 or G^2 with $df = (I - 1)(J - 1)$; falling farther out in the tail produces a smaller P -value. The potential discrepancy in power increases as I and J increase.[5](#)

3.4.5 Sensitivity to Choice of Scores

Often, it is unclear how to assign scores to statistics that require them, such as M^2 in Section 3.4.1. Cochran (1954) noted that “any set of scores gives a *valid* test, provided that they are constructed without consulting the results of the experiment. If the set of scores is poor, in that it badly distorts a numerical scale that really does underlie the ordered classification, the test will not be sensitive. The scores should therefore embody the best insight available about the way in which the classification was constructed and used.” Ideally, the scale is chosen by a consensus of experts, and subsequent interpretations use that same scale.

How sensitive are analyses to the choice of scores? There is no simple answer.⁶ For most data sets, different choices of monotone scores give similar results. Scores that are linear transforms of each other, such as (1,2,3,4) and (0,2,4,6), have the same absolute correlation and hence the same M^2 . Results *may* depend on the scores, however, when the data are highly unbalanced, with some categories having many more observations than others.

3.4.6 Example: Infant Birth Defects by Maternal Alcohol Consumption

Graubard and Korn (1987) used [Table 3.8](#) to illustrate the potential dependence. It refers to a prospective study of maternal drinking and birth defects. After the first three months of pregnancy, the women in the sample completed a questionnaire about alcohol consumption. Following childbirth, observations were recorded on the presence or absence of congenital sex organ malformations. When a variable is nominal but has only two categories, statistics that treat it as ordinal are still valid. For instance, we can artificially regard malformation as ordinal, treating “present” as “high” and “absent” as “low.” With only two rows, any set of distinct row scores is a linear transformation of any other set and gives the same M^2 value. Alcohol consumption, measured as the average number of drinks per day, is an ordinal explanatory variable. This groups a naturally continuous variable, and we first use the scores $\{v_1 = 0, v_2 = 0.5, v_3 = 1.5, v_4 = 4.0, v_5 = 7.0\}$, the last score being somewhat arbitrary. For this choice, $M^2 = 6.57$, for which the P -value is 0.010. By contrast, for the equally spaced row scores (1,2,3,4,5), $M^2 = 1.83$, giving a much weaker conclusion ($P = 0.18$).

Table 3.8 Data for Which Test Results Depend Greatly on Scores for Alcohol Consumption

Malformation	Alcohol Consumption (average number of drinks per day)				
	0	< 1	1–2	3–5	≥ 6
Absent	17,066	14,464	788	126	37
Present	48	38	5	1	1

Source: Reprinted with permission from the Biometric Society (Graubard and Korn 1987).

An alternative approach uses the data to form the scores automatically, with ranks as the category scores. All subjects in a category receive the average of the ranks that would apply for a complete ranking of the sample from 1 to n . These are called *midranks*. When X and Y are both ordinal and M^2 uses midrank scores, the correlation on which M^2 is based is called *Spearman's rho*. For [Table 3.8](#), the 17,114 subjects at level 0 for alcohol consumption share ranks 1 through 17,114. Each receives the average of these ranks, which is the midrank $(1 + 17,114)/2 = 8557.5$. Similarly, the midranks for the last four categories are 24,365.5, 32,013, 32,473, and 32,555.5. These scores yield $M^2 = 0.35$ and a weaker conclusion yet ($P = 0.55$).

Why does this happen? Adjacent categories having relatively few observations necessarily have similar midranks. The midranks are similar for the final three categories, since those categories have few observations compared with the first two categories. This scoring scheme treats alcohol consumption level 1-2 drinks (category 3) as much closer to consumption level ≥ 6 drinks (category 5) than to consumption level 0 drinks (category 1).

This seems inappropriate. It is usually better to select scores that reflect perceived distances between categories. When uncertain about this choice, a sensitivity analysis should be performed, selecting two or three sensible choices and checking whether results are similar. Equally spaced scores often provide a reasonable compromise when the category labels do not suggest obvious choices, such as the categories (liberal, moderate, conservative) for political philosophy.

3.4.7 Trend Tests for $I \times 2$ and $2 \times J$ Tables

When I or J equals 2, the tests based on linear or monotonic trend simplify to well-established procedures. With binary X , $2 \times J$ tables occur in comparisons of two groups, such as when the rows represent two treatments. Using scores $\{u_1 = 0, u_2 = 1\}$ for levels of X , the covariation measure $\sum_j u_i v_j p_{ij}$ in M^2 simplifies to $\sum_j v_j p_{2j}$. Divided by the proportion of subjects in row 2, it gives the mean score for that row. In fact, M^2 is then directed toward detecting differences between the two row means of the scores on Y .

With midrank scores for Y , the test using M^2 for $2 \times J$ tables is sensitive to differences in mean ranks for the two rows. This test is called the *Wilcoxon* or *Mann-Whitney test*. The large-sample version of that test uses a standard normal z statistic that is equivalent to $z = (C - D)/SE_0$ based on the difference between the numbers of concordant and discordant pairs relative to the null SE. The square of the statistic is equivalent to M^2 , using arbitrary row scores and midranks for the columns. For summarizing the difference between the two groups, related measures such as $\Delta = P(Y_1 > Y_2) - P(Y_2 > Y_1)$ are also relevant (Section 2.4.6). Ryu and Agresti (2008) proposed score-type confidence intervals for such measures.

When Y has two levels, the table has size $I \times 2$. The linear trend statistic then refers to a linear trend in the probability of either response category, such as the probability of malformation as a function of alcohol consumption. The test in that case, often called the *Cochran-Armitage trend test*, is presented in Section 5.3.5.

3.4.8 Nominal-Ordinal Tables

Inference using measures such as the correlation and gamma is appropriate when both classifications are ordinal. When one is nominal with more than two categories, other statistics are needed. One is based on summarizing the variation among means on the ordinal variable in the various categories of the nominal variable. We defer discussion of this case to Note 3.7, Exercise 3.37, and Section 8.4.3.

3.5 SMALL-SAMPLE INFERENCE FOR CONTINGENCY TABLES

The inferential methods of the preceding four sections are large-sample methods. When n is small, alternative methods use *exact* small-sample distributions rather than large-sample approximations. In this section we describe small-sample tests of independence, starting with one that R. A. Fisher proposed for 2×2 tables.

3.5.1 Fisher's Exact Test for 2×2 Tables

In Section 16.5.1 we show that, under H_0 : independence, conditioning on the marginal totals of the contingency table produces a null distribution for the cell counts that does not depend on unknown parameters. Usually both margins are not naturally fixed. For Poisson sampling nothing is fixed, for multinomial sampling only n is fixed, and for independent binomial row samples only the row marginal totals are fixed. In any of these cases, conditioning on both sets of marginal totals in a 2×2 table yields the *hypergeometric distribution*

$$(3.17) \quad p(t) = P(n_{11} = t) = \frac{\binom{n_{1+}}{t} \binom{n_{2+}}{n_{+1}-t}}{\binom{n}{n_{+1}}}.$$

This formula expresses the distribution of $\{n_{ij}\}$ in terms of only n_{11} . Given the marginal totals, n_{11} determines the other three cell counts. The range of possible values for n_{11} is $m_- \leq n_{11} \leq m_+$, where $m_- = \max(0, n_{1+} + n_{+1} - n)$ and $m_+ = \min(n_{1+}, n_{+1})$.

For 2×2 tables, independence is equivalent to the odds ratio $\theta = 1$. To test $H_0: \theta = 1$, the P -value is the sum of certain hypergeometric probabilities. To illustrate, consider $H_a: \theta > 1$. For the given marginal totals, tables having larger n_{11} have larger sample odds ratios and hence stronger evidence in favor of H_a . Thus, the P -value equals $P(n_{11} \geq t_0)$, where t_0 denotes the observed value of n_{11} . This test for 2×2 tables is called *Fisher's exact test*.

3.5.2 Example: Fisher's Tea Drinker

R. A. Fisher (1935a) described the following experiment from his days working at Rothamsted Experimental Station, an agriculture research lab north of London. Dr. Muriel Bristol, a colleague of Fisher's, claimed that when drinking tea she could distinguish whether milk or tea was added to the cup first (she preferred milk first). To test her claim, Fisher asked her to taste eight cups of tea, four of which had milk added first and four of which had tea added first. She knew there were four cups of each type and had to predict which four had the milk added first. The order of presenting the cups to her was randomized.

[Table 3.9](#) shows a possible result. Distinguishing the order of pouring better than with pure guessing corresponds to $\theta > 1$, reflecting a positive association between order of pouring and the prediction. We conduct Fisher's exact test of $H_0: \theta = 1$ against $H_a: \theta > 1$.

Table 3.9 Data for Fisher's Tea-Tasting Experiment

Poured First	Guess Poured First		Total
	Milk	Tea	
Milk	3	1	4
Tea	1	3	4
Total	4	4	

Source: Based on experiment described by Fisher (1935a).

The experimental design fixed both marginal distributions, since Dr. Bristol had to predict which four cups had milk added first. Thus, the hypergeometric applies naturally for the null distribution of n_{11} . The P -value for Fisher's exact test is the null probability of [Table 3.9](#) and of tables having even more evidence in favor of her claim. The observed table, $t_0 = 3$ correct choices of the cups having milk added first, has null probability

$$\binom{4}{3} \binom{4}{1} / \binom{8}{4} = 0.229.$$

The only table that is more extreme in the direction of H_a has $n_{11} = 4$ correct. It has a probability of 0.014. The P -value is $P(n_{11} \geq 3) = 0.243$. This result does not establish an association between the actual order of pouring and her predictions. According to Fisher's daughter (Box 1978, p. 134), in reality Dr. Bristol did convince Fisher of her ability.

3.5.3 Two-Sided P -Values for Fisher's Exact Test

For the one-sided alternative, the same P -value results using tables ordered according to larger n_{11} , larger odds ratio, or larger difference of proportions (Davis 1986a). For the two-sided alternative, different criteria can have different P -values.

For a two-sided P -value, the most common approach (Irwin 1935) sums $P(n_{11} = t)$ in (3.17) for counts t such that $p(t) \leq p(t_0)$; that is, the P -value is $P[p(n_{11}) \leq p(t_0)]$ for the observed value t_0 . Another possibility sums $p(t)$ for tables that are farther from H_0 ; that is,

$$P\text{-value} = P[|n_{11} - E(n_{11})| \geq |t_0 - E(n_{11})|],$$

where the hypergeometric $E(n_{11}) = n_{1+} n_{+1}/n$. This is identical to $P(X^2 \geq X^2_{t_0})$ for observed Pearson statistic $X^2_{t_0}$. A third approach doubles the minimum one-sided P -value, that is, $2 \min[P(n_{11} \geq t_0), P(n_{11} \leq t_0)]$, but this can exceed 1. A fourth approach (Blaker 2000) uses $Q = \min[P(n_{11} \geq t_0), P(n_{11} \leq t_0)]$ plus an attainable probability in the other tail that is as close as possible to, but not greater than, that one-tailed probability. This P -value can be expressed as $P(Q \leq q_0)$ for observed value q_0 of Q .

Each approach has advantages and disadvantages (3.10). They can provide different results because of the discreteness and potential skewness. The approach of ordering tables by a distance measure from H_0 , such as X^2 , extends naturally to $I \times J$ tables. Exact tests for that more general case are deferred to Section 16.5.2.

In practice, two-sided tests are more common than one-sided. Partly this is so that researchers can avoid charges of bias in giving evidence that supports their predicted direction for an effect. To conduct a test of size 0.05 when you truly believe that the effect has a particular direction, you can conduct the one-sided test at the 0.025 level to guard against criticism. For instance, in the 1998 document *Biostatistical Principles for Clinical Trials*, the International Conference on Harmonization (ICH E9) stated: “The approach of setting type I errors for one-sided tests at half the conventional type I error used in two-sided tests is preferable in regulatory settings. This promotes consistency with two-sided confidence intervals that are generally appropriate for estimating the possible size of the difference between two treatments.”

3.5.4 Confidence Intervals Based on Conditional Likelihood

Small-sample methods also apply to estimation. An approach discussed in Section 7.3 uses a *conditional likelihood function* to eliminate nuisance parameters by conditioning on their sufficient statistics. This is useful even with large-sample confidence intervals. In fact, the intervals for the odds ratio in Sections 3.2.5 and 3.2.6 that utilize joint distributions with fixed column margins as well as row margins are actually *conditional* score and profile likelihood confidence intervals.

Small-sample interval estimation entails other complicating issues resulting from discreteness and evaluating performance over an entire parameter space. We defer discussion of those methods to Sections 16.6.4 and 16.6.8.

3.5.5 Discreteness and Conservatism Issues

The hypergeometric distribution (3.17) is highly discrete for small samples, because n_{11} and hence the P -value can assume relatively few values. It is usually not possible to achieve a fixed significance level (size) such as 0.05.

In the tea-tasting experiment, for instance, n_{11} can equal only 4,3,2,1,0. The one-sided P -values are restricted to 0.014, 0.243, 0.757, 0.986, and 1.0. If we reject H_0 when the P -value ≤ 0.05 , then 0.05 is not the probability of type I error. Only the P -value of 0.014 does not exceed 0.05; thus, when H_0 is true, the probability of falsely rejecting it is 0.014, not 0.05. In this sense, the traditional approach to hypothesis testing is conservative: The true probability of type I error is bounded above by the nominal level.

It is possible to achieve any fixed significance level by data-unrelated randomization on the boundary of the critical region, in deciding whether to reject H_0 . For the tea-tasting experiment, suppose that we reject H_0 when $n_{11} = 4$, we reject H_0 with probability 0.157 when $n_{11} = 3$, and we do not reject H_0 otherwise; that is, when $n_{11} = 2$, we generate a uniform random variable U over [0, 1] and reject H_0 if $U < 0.157$. For expectation taken with respect to the null hypergeometric distribution of n_{11} , the significance level then equals

$$\begin{aligned} P(\text{reject } H_0) &= E[P(\text{reject } H_0|n_{11})] \\ &= 1.0(0.014) + 0.157(0.229) + 0.0 \times P(n_{11} \leq 2) = 0.05. \end{aligned}$$

With the randomization extension, Tocher (1950) showed that Fisher's test is uniformly most powerful unbiased (UMPU) for the chosen size (here, 0.05). This property follows from conditioning on a sufficient statistic that is complete and has distribution in the exponential family (Lehmann and Romano 2005, Sec. 4.4-4.7).

In practice, randomization having nothing to do with the data is unacceptable, and sensible tests for this hypothesis are biased (Exercise 3.43). We recommend simply reporting the P -value. To reduce conservativeness, report the mid P -value (Section 1.4.4). The test is no longer guaranteed to have true P (type I error) no greater than the nominal value, but in practice it is rarely much greater. For the one-sided test with the tea-tasting data,

$$\text{mid } P\text{-value} = \left(\frac{1}{2}\right) P(n_{11} = 3) + P(n_{11} > 3) = 0.129.$$

For a two-sided test using a criterion such as X^2 , we would add half the probability of the observed X^2 value to the probabilities for larger values, for tables with the given margins.

3.5.6 Small-Sample Unconditional Tests of Independence

A common sampling assumption for analyses comparing two groups on a binary response is that the rows are independent binomial samples. Then, only $\{n_{i+}\}$ are naturally fixed. For Poisson and multinomial sampling schemes, neither marginal distribution is fixed. For such cases it may seem artificial to condition on *both* sets of marginal counts. An alternative small-sample test, designed for independent binomial samples, conditions on only the row totals.

Under binomial sampling with parameter π_i in row i , consider testing $H_0: \pi_1 = \pi_2$ using some test statistic T , such as the Pearson X^2 . For fixed $\{n_{i+}\}$, T can take a discrete set of values, one of which is the observed value t_0 . Given $\pi_1 = \pi_2 = \pi$, the P -value is $P_\pi(T \geq t_0)$, calculated using the product of the two binomial probability mass functions. This is the sum of the product binomial probabilities for those pairs of binomial samples that have $T \geq t_0$. Since π is unknown, the actual P -value is defined as

$$(3.18) \quad P = \sup_{0 \leq \pi \leq 1} P_\pi(T \geq t_0).$$

This is an *unconditional* small-sample test of independence. Like Fisher's exact test, the true size is no greater than the nominal value [e.g., if we reject when $P \leq 0.05$, the actual P (type I error) ≤ 0.05].

We illustrate using test statistic X^2 for the 2×2 table having entries (3,0/0,3), by row, with fixed row totals (3,3) as binomial sample sizes. The sample $X^2 = 6.0$. This X^2 value for the observed table and for table (0,3/3,0) is the maximum possible. For a given value π for $\pi_1 = \pi_2$, the probability of the first table is $[\pi^3(1 - \pi)^0][\pi^0(1 - \pi)^3] = \pi^3(1 - \pi)^3$ (3 successes and 0 failures in the first row and 0 successes and 3 failures in the second), the product of two binomial probabilities. Similarly, the probability of the second table is $(1 - \pi)^3\pi^3$. Thus, the conditional P -value is $P_\pi(X^2 \geq 6) = 2\pi^3(1 - \pi)^3$, the sum of the product binomial probabilities for those two tables. The supremum of this over $0 \leq \pi \leq 1$ occurs at $\pi = \frac{1}{2}$, giving overall P -value equal to $2(0.5)^3(0.5)^3 = 0.031$. By contrast, the two-sided Fisher's exact test has P -value equal to $2 \binom{3}{0} \binom{3}{3} / \binom{6}{3} = 0.100$.

Barnard (1945,1947) was the first to propose an unconditional test comparing binomial parameters, although he later (1949) withdrew it in favor of Fisher's exact test. His method forms a critical region by adding points according to certain criteria until the supremum over π of the probability of points in the region is as close to the desired size as possible. Several authors have since proposed related tests. Suissa and Shuster (1985) used (3.18) with T as the pooled or unpooled z statistic for comparing two proportions, which give identical P -values when $n_{1+} = n_{2+}$. Haber (1986) also used the pooled statistic. Boschloo (1970) suggested using an increased significance level for the conditional test considered for all the possible response marginal distributions, such that the unconditional size at each value of π under H_0 (averaged over the possible marginal distributions) is no greater than the nominal level. This essentially uses the P -value from Fisher's exact test as a test statistic (Mehrotra et al. 2003, Lin and Yang 2009). That is, the P -value for Boschloo's test is the supremum over π of the product binomial probability that the Fisher's P -value is less than or equal to the observed Fisher P -value. The Boschloo test is necessarily at least as powerful as Fisher's test since its rejection region contains that for Fisher's test.

3.5.7 Conditional Versus Unconditional Tests

Since Barnard introduced the unconditional test, many statisticians have debated the proper way to conduct small-sample analyses of 2×2 tables. Fisher criticized the unconditional approach, arguing that possible samples with quite different numbers of successes than observed were not relevant. In Fisher's (1945) view, "... the existence of these less informative possibilities should not affect our judgment of significance based on the series actually observed.... The fact that such an unhelpful outcome as these might occur ... is surely no reason for enhancing our judgment of significance in cases where it has not occurred;... it is only the sampling distribution of samples of the same type that can supply a rational test of significance."

An adaptation of the unconditional approach by Berger and Boos (1994) addresses this criticism somewhat. They took the supremum for the P -value over a confidence interval of values for the nuisance parameter π rather than over all possible values. Their unconditional P -value is

$$P = \sup_{\pi \in C_\gamma} P_\pi(T \geq t_0) + \gamma,$$

where C_γ is a confidence interval for π with coverage probability at least $(1 - \gamma)\%$. Here, γ is taken to be very small (e.g., 0.001), and the test maintains the guaranteed upper bound on size.

Other arguments in favor of conditioning on both sets of marginal totals are that the conditional approach provides a simple way to eliminate nuisance parameters that generalizes to many other contingency table problems, and the margins contain little information about the association (see Note 3.9). In an informal sense, the margins can contain much more information about the range of plausible values for the difference of proportions than the odds ratio, as illustrated by a 2×2 comparison of two groups of size 50 each when the total sample contains only 1 success. Arguments against conditioning partly concern the increased discreteness that occurs. The few possible values for n_{11} make it difficult to obtain a small P -value. In repeated use with a nominal significance level, the actual type I error probability may be much smaller than the nominal value and the power may suffer. Finally, for inference about nonnull values (e.g., confidence intervals), we will see (Section 16.6) that the conditional approach applies for the odds ratio but not for other association measures.

The conservatism issue is partly unavoidable. Statistics having discrete distributions are necessarily conservative in terms of achieving nominal significance levels. Because an unconditional test fixes only one margin, however, it has many more tables in the reference set for its sampling distribution. That distribution is less discrete, and a richer array of possible P -values occurs than with Fisher's exact test. An unconditional test tends to be less conservative and more powerful than Fisher's exact test. A disadvantage is that computations are very intensive for more complex problems, such as tables larger in size than 2×2 .

If a table truly has two independent binomial samples, the unconditional approach seems sensible. See Kempthorne (1979) for a cogent argument. The conditional approach is useful for other cases, such as with convenience samples in experiments. In a randomized clinical trial, a sample of n available subjects is randomly allocated to two treatments. The samples are not binomials, as they are not random samples from two populations of interest. We could focus on the sample alone and consider the probability of a result at least as extreme as observed if there truly is no treatment effect. For instance, out of all possible ways of choosing n_{1+} of the n subjects for treatment 1, for what proportion of samples would n_{11} be at least as large as observed? Under the null hypothesis of no treatment effect, the same overall response distribution (n_{+1}, n_{+2}) of successes and failures occurs regardless of the allocation of subjects to treatments. Thus, the column margin is also naturally fixed. This argument leads to hypergeometric null probabilities and Fisher's exact test (Greenland 1991) or its mid- P adaptation. This randomization-based approach does not extend, however, to nonnull effect values and hence to confidence intervals.

Sometimes both sets of marginal totals are naturally fixed, such as in [Table 3.9](#). Then, the high degree of discreteness is unavoidable and it is natural to use Fisher's exact test and its mid- P adaptation.

3.6 BAYESIAN INFERENCE FOR TWO-WAY CONTINGENCY TABLES

Bayesian methods are relatively straightforward for estimating cell probabilities or association measures for contingency tables. When we distinguish between response and explanatory variables, we treat the cell counts as realizations of independent multinomial samples and formulate prior distributions (such as Dirichlet distributions) for the multinomial parameters. When both variables are response variables, we can treat the cell counts as a single multinomial sample and formulate a prior distribution for the entire set of cell probabilities.

3.6.1 Prior Distributions for Comparing Proportions in 2×2 Tables

We consider first the comparison of parameters for two independent binomial samples summarized in a 2×2 contingency table. Let Y_i denote a binomial $\text{bin}(n_i, \pi_i)$ variate, $i = 1, 2$.

The conjugate Bayesian approach uses independent $\text{beta}(\alpha_{i1}, \alpha_{i2})$ prior densities for $\pi_i = 1, 2$. This yields independent posterior $\text{beta}(y_i + \alpha_{i1}, n - y_i + \alpha_{i2})$ densities for π_i , $i = 1, 2$.

With independent continuous prior densities, the prior and posterior probability of homogeneity, $\pi_1 = \pi_2$, is zero. To allow $P(\pi_1 = \pi_2 | y_1, n_1; y_2, n_2) > 0$, we could use a prior distribution that has a positive $P(\pi_1 = \pi_2)$. For example, we could use a prior distribution for which π_1 and π_2 have $\text{beta}(\alpha_1, \alpha_2)$ distributions, such that $\pi_1 = \pi_2$ with probability γ and π_1 is independent of π_2 with probability $1 - \gamma$.

Even if we use a model having $P(\pi_1 = \pi_2) = 0$, in practice, it is possible to treat π_1 and π_2 as dependent, a priori. For instance, if we knew that $\pi_1 = 0.02$, then in many applications conditionally this would induce the subjective belief that π_2 is also close to 0. Howard (1998) suggested a prior distribution for correlated (π_1, π_2) . For odds ratio θ , he amended the independent beta prior distributions by using prior density function proportional to

$$e^{-(1/2\sigma^2)[\log(\theta)]^2} \pi_1^{\alpha_{11}-1} (1-\pi_1)^{\alpha_{12}-1} \pi_2^{\alpha_{21}-1} (1-\pi_2)^{\alpha_{22}-1}.$$

The correlation decreases as σ increases, with independent beta densities resulting in the limit as $\sigma \rightarrow \infty$. For the amended Jeffreys prior distribution ($\alpha_{i1} = \alpha_{i2} = 0.50$), when $\sigma = (1, 2, 3)$ the correlations are (0.84, 0.59, 0.41) between π_1 and π_2 (Agresti and Min 2005a).

An alternative prior distribution that can incorporate correlation is a bivariate normal distribution for $[\text{logit}(\pi_1), \text{logit}(\pi_2)]$. Taking marginal means of 0 and standard deviations of about 3 is relatively uninformative. In that case, when $\text{corr}[\text{logit}(\pi_1), \text{logit}(\pi_2)] = 0.50$, $\text{corr}(\pi_1, \pi_2) = 0.45$. This case gives similar results as Howard's prior with $\sigma = 3$. An alternative way of inducing dependence is to use a hierarchical prior, but that is beyond our scope here.

3.6.2 Posterior Probabilities Comparing Proportions

For the conjugate prior structure of independent beta densities, independent $\text{beta}(y_i + \alpha_{i1}, n - y_i + \alpha_{i2})$ posterior densities determine inference about π_i , $i = 1, 2$. We can evaluate through simulation or numerical integration relevant posterior probabilities such as $P(\pi_1 < \pi_2 | y_1, n_1; y_2, n_2)$ and $P(\pi_1 > \pi_2 | y_1, n_1; y_2, n_2)$ and construct posterior intervals for summary measures such as $\pi_1 - \pi_2$ and the odds ratio.

In many applications, one group (say, group 1) receives a new treatment and group 2 receives some standard treatment or placebo, and the purpose of the study is to analyze whether the response tends to be better with the new treatment. Then, in terms of “success” probabilities, we could regard $\pi_1 \leq \pi_2$ as a null condition and $\pi_1 > \pi_2$ as an alternative. Then, $P(\pi_1 \leq \pi_2 | y_1, n_1; y_2, n_2)$ is a sort of Bayesian P -value. Howard (1998) showed that with use of Jeffreys priors with $\alpha_{i1} = \alpha_{i2} = 0.5$, $P(\pi_1 \leq \pi_2 | y_1, n_1; y_2, n_2)$ approximately equals the one-sided P -value for the large-sample z test (3.12) (which is the signed square root of the Pearson X^2 statistic) for testing $H_0 : \pi_1 = \pi_2$ against $H_a : \pi_1 > \pi_2$.

For testing $H_0 : \pi_1 \leq \pi_2$ against $\pi_1 > \pi_2$, Altham (1969) showed how the posterior $P(\pi_1 \leq \pi_2 | y_1, n_1; y_2, n_2)$ relates to the one-sided P -value for Fisher’s exact test. With prior beta hyperparameters $\alpha_{i1} = \alpha_{i2} = \gamma$, $i = 1, 2$, and $0 \leq \gamma \leq 1$, Altham showed that $P(\pi_1 \leq \pi_2 | y_1, n_1; y_2, n_2)$ is smaller than the Fisher P -value by no more than the null probability of the observed data. They are identical when we use the improper prior densities with $(\alpha_{11}, \alpha_{12}) = (1, 0)$ and $(\alpha_{21}, \alpha_{22}) = (0, 1)$. These priors favor H_0 , in effect penalizing against concluding that $\pi_1 > \pi_2$. For example, for any data having the same sample proportion of successes $0 < p < 1$ for the two groups, $P(\pi_1 \leq \pi_2 | y_1, n_1; y_2, n_2) > 0.50$. So, Fisher’s one-sided exact P -value corresponds to a Bayesian inference with conservative prior distributions.

3.6.3 Posterior Intervals for Association Parameters

We could also use the Bayesian approach to construct posterior intervals for the difference of proportions, relative risk, and odds ratio. Any particular prior distribution for (π_1, π_2) induces corresponding prior distributions for the measures themselves. For instance, with independent uniform prior distributions for π_1 and π_2 , $\pi_1 - \pi_2$ has a triangular density over $(-1, +1)$, and the log odds ratio has the Laplace density (Nurminen and Mutanen 1987).

Similarly, a posterior distribution for (π_1, π_2) induces posterior distributions for the measures. For the case of independent beta posterior distributions for π_1 and π_2 , we can easily simulate the posterior distribution of a measure of association by simulating values from the beta distributions. Thus, it is easy to simulate reasonable approximations for posterior intervals, for instance, forming the 95% equal-tail interval using values between the simulated 2.5 percentile and 97.5 percentile of the posterior distribution.

Finding more precise intervals requires better approximations. Let $F_\omega(t)$ denote the cdf for the posterior distribution of a generic measure of association ω . In terms of independent beta posterior densities $f(\pi_i | y_i, n_i)$ for π_i , $i = 1, 2$,

$$F_\omega(t) = \int \int_{S_t} f(\pi_1 | y_1, n_1) f(\pi_2 | y_2, n_2) d\pi_1 d\pi_2,$$

where $S_t = \{(\pi_1, \pi_2) : \omega \leq t, 0 < \pi_1, \pi_2 < 1\}$. The equal-tail 95% posterior interval \underline{L} (\bar{U}) satisfies $F_\omega(\underline{L}) = 0.025$ and $F_\omega(\bar{U}) = 0.975$. Hashemi et al. (1997) and Nurminen and Mutanen (1987) gave integral expressions for the posterior distributions of the difference of proportions, relative risk, and odds ratio.

To obtain good frequentist performance in terms of maintaining coverage probability close to the nominal level over the entire parameter space, it is best to use quite diffuse priors. Even uniform priors are often too informative. Agresti and Min (2005a) recommended, in agreement with Brown et al. (2001) in the single binomial case, using independent Jeffreys priors for π_1 and π_2 .

3.6.4 Example: Urn Sampling Gives Highly Unbalanced Treatment Allocation

For small samples, results can depend strongly on the choice of prior distribution, as shown by an example from a clinical trial discussed by Begg (1990). For an urn sampling method to allocate patients to treatments, the 11 patients allocated to the experimental treatment were all successes and the only patient allocated to the control treatment was a failure. That is, the table has rows (11,0) and (0,1).

The 95% equal-tail posterior interval for the odds ratio is (1.2, 218.4) for independent beta(2, 2) priors, (1.7, 4677) for uniform (beta(1, 1)) priors, and (3.3, 1.4×10^6) for Jeffreys (beta(0.5, 0.5)) priors. By contrast, the frequentist 95% confidence interval based on inverting the large-sample score test is (4.5, ∞). Incorporating prior beliefs with a mean of no effect causes the lower bound for the odds ratio to be pulled considerably toward the no effect value of 1.0.

With uniform priors, the posterior densities are beta(12, 1) for π_1 and beta(1, 2) for π_2 . A simple way to estimate precisely the posterior $P(\pi_1 \geq \pi_2 | y_1, n_1; y_2, n_2)$ is to generate a huge number of beta random variables from these two densities and observe the proportion of cases for which $\pi_1 \geq \pi_2$. We then find that $P(\pi_1 \geq \pi_2 | y_1, n_1; y_2, n_2) = 0.99$. There is strong evidence that the experimental treatment is better than control.

3.6.5 Highest Posterior Density Intervals

An alternative approach to constructing posterior intervals uses highest posterior density (HPD) intervals. When applied to the odds ratio and relative risk, this method has a serious disadvantage: It is not invariant under nonlinear parameter transformation. Specifically, suppose (L, U) is a 95% HPD interval based on the posterior distribution of the odds ratio θ . Then, the 95% HPD interval based on the posterior distribution of $1/\theta$, which is relevant if we reverse the labeling of the two groups being compared, is not $(1/U, 1/L)$. In fact, it can be considerably different. This happens because the 95% region of highest density for a random variable X is not the inverse mapping of the 95% region of highest density for $1/X$.

To illustrate, consider uniform prior densities for π_1 and π_2 when $n_1 = n_2 = 10$. When $y_1 = 1$ and $y_2 = 5$, $\theta = 1/9$ and the Bayes 95% HPD interval for θ is $(0.0006, 0.82)$; when $y_1 = 5$ and $y_2 = 1$, $\theta = 9$ and the Bayes 95% HPD interval is $(0.17, 38.23)$, which is very different from $(1/0.82, 1/0.0006)$. By contrast, the 95% equal-tail confidence intervals for the odds ratios with uniform priors are $(0.017, 1.10)$ when $y_1 = 1$ and $y_2 = 5$ and $(0.91 = 1/1.10, 57.9 = 1/0.017)$ when $y_1 = 5$ and $y_2 = 1$. In another example, for tables with $\theta = 0$, the HPD interval with diffuse priors is typically of the form $(0, U)$, but when rows are interchanged so that $\theta = \infty$, the HPD interval has a finite upper bound (Agresti and Min 2005a).

HPD invariance to group labeling does occur on the log scale for the odds ratio and relative risk, because the relevant parameter is a difference (e.g., log odds ratio = difference of log odds) and so is linearly rather than nonlinearly transformed by a relabeling of the groups. However, users interpret the magnitude of the odds ratio on its original scale rather than the log scale. So, the lack of invariance when constructing HPD intervals on the original scale is to us a compelling reason not to use the HPD approach for the odds ratio or relative risk.

An exception when the HPD interval seems sensible is when the posterior density is monotone. Then, excluding both upper and lower tails of that distribution with the equal-tail method seems inappropriate. For example, suppose the sample odds ratio is 0 and the HPD interval has form $(0, U)$, with the two binomials relabeled if necessary so this is the case. The HPD interval then seems more relevant than the equal-tail interval. However, then it seems sensible when groups or outcome categories are interchanged to use the corresponding posterior interval $(1/U, \infty)$, which is not HPD.

Consider the difference of proportions. When $\pi_1 - \pi_2$ takes its boundary values of $+1$ or -1 , the posterior density is monotone with the Jeffreys prior or more diffuse priors, and close to monotone for priors that are more informative than the Jeffreys prior (Agresti and Min 2005a). So, the HPD interval for $\pi_1 - \pi_2$ seems sensible. With the Jeffreys prior, the HPD interval then has the form $(L, 1)$ or $(-1, U)$.

3.6.6 Testing Independence

For 2×2 tables, Bayesian inference about whether two binary variables are independent can be based directly on posterior tail probabilities and intervals for association parameters, such as we illustrated in Section 3.6.4. For $I \times J$ tables, such inference is not as straightforward, because independence relates to $(I - 1)(J - 1)$ parameters instead of a single parameter.

One approach for $I \times J$ tables forms a Bayes factor that is a ratio comparing the probability of the data under (1) H_0 : independence and (2) H_a : association (see Note 3.11). Converting this Bayes factor to the posterior probability that H_0 is true requires choosing a prior probability that H_0 is true. Naturally, the posterior probability is highly dependent on the choice of this prior probability. Gunel and Dickey (1974) considered independence in two-way contingency tables under the usual sampling models. Conjugate gamma priors for the Poisson model induce priors in each further conditioned model. They showed that the Bayes factor for independence itself factors, highlighting the evidence residing in the marginal totals.

Ultimately, it is more informative to focus on estimating parameters that describe the association. With two ordinal response variables, we could summarize the evidence about a positive or negative association as summarized by a measure such as the correlation or gamma. For example, using the approach of Section 1.6.3 of combining a Dirichlet prior distribution for cell probabilities with a multinomial likelihood function yields a Dirichlet posterior distribution for the cell probabilities. This induces a posterior distribution for the ordinal measure of interest, yielding a posterior interval and posterior probabilities of a positive association and of a negative association.

3.6.7 Empirical Bayes and Hierarchical Bayesian Approaches

Some methodologists find it appealing to treat parameters as random variables having distributions but dislike the subjectivity inherent in the Bayesian approach from selecting a prior distribution. An alternative way of implementing a Bayesian approach is to let the data suggest hyperparameter values for use in the prior distribution. This is called the *empirical Bayes* approach. Most commonly, this approach uses the prior hyperparameter values that maximize the marginal probability of the observed data, integrating out the parameters with respect to that prior (e.g., Efron and Morris 1975). A related approach estimates the prior that has Bayes estimator with smallest total mean squared error (Exercise 3.46). I. J. Good seems to have first used an empirical Bayesian approach with contingency tables, estimating hyperparameters in gamma and log-normal priors for association factors. Good (1965) used it to estimate the hyperparameter value for a symmetric Dirichlet prior for multinomial parameters.

A disadvantage of the empirical Bayesian approach is not accounting for the source of variability due to substituting estimates for prior hyperparameters. An alternative approach not having this disadvantage is *hierarchical Bayes*, which lets the prior hyperparameters themselves have a second-stage prior distribution. For multinomial data, for example, Good (1965,1976) noted that Dirichlet priors do not always provide sufficient flexibility. He proposed a hierarchical approach of specifying distributions for the Dirichlet hyperparameters, treating $\{\alpha_i\}$ in the Dirichlet prior as unknown and specifying a second-stage prior for them. This approach gains greater generality at the expense of giving up the simple conjugate Dirichlet form for the posterior.

Most of the empirical Bayes and hierarchical Bayes literature refers to estimating multiple parameters, such as several binomial parameters $\{\pi_i\}$. For instance, at stage 1, given μ and σ , we might assume that $\{\text{logit}(\pi_i)\}$ are independent from a $N(\mu, \sigma^2)$ distribution, and at stage 2 assume a highly disperse normal prior for μ and an inverse chi-squared prior distribution for σ^2 . Leonard (1972) proposed an approach of this type, for which the posterior mean estimate of $\text{logit}(\pi_i)$ is approximately a weighted average of $\text{logit}(p_i)$ and $\{\text{logit}(p_j), j \neq i\}$.

3.7 EXTENSIONS FOR MULTIWAY TABLES AND NONTABULATED RESPONSES

The methods of this chapter extend to multiway contingency tables. For instance, tests of independence for two-way tables extend to tests of conditional independence in three-way tables. In future chapters we present such methods with models that provide a basis for defining relevant parameters and their statistical inferences.

3.7.1 Categorical Data Need Not Be Contingency Tables

Examples so far have presented categorical data in the format of contingency tables. However, this book has broader focus than contingency table analysis. Models for categorical response variables can have continuous as well as discrete explanatory variables. Even when all or most variables are categorical, source data files are not usually contingency tables but have the form of a line of data for each subject. The first three lines in a data file containing responses of a survey of subjects measuring gender, race, education (1 = less than high school, 2 = high school or some college, 3 = college graduate), and attitude toward homosexuality (1 = tolerant, 2 = homophobic) might be:

Subject	Gender	Race	Education	Attitude
1	f	w	2	1
2	m	b	3	1
3	m	w	1	2

Software can read data files of this type and then conduct analyses that may or may not involve forming contingency tables.

In the next chapter we introduce the modeling framework used in the rest of the book. All the methods that we've studied in this chapter result from inferences for parameters in simple versions of these models.

NOTES

Section 3.1: Confidence Intervals for Association Parameters

3.1 Standard errors: Goodman and Kruskal (1963, 1972) provided standard errors for many association measures and extended (3.9) for independent multinomial sampling. For adaptations of the Wald interval (3.2) for $\log \theta$ that better handle zero cell counts, see Agresti (1999) and Gart (1971). Agresti and Caffo (2000) showed that as in the single-sample case (Exercise 1.25), the Wald interval (3.4) for $\pi_1 - \pi_2$ behaves much better after adding two pseudo-observations of each type (one of each type in each sample). Fagerland et al. (2012) compared various confidence interval methods for the difference of proportions, odds ratio, and relative risk.

3.2 Multiple comparisons: Agresti et al. (2008) proposed a method for multiple comparisons using effect measures for comparing proportions for g groups that is an analog of Tukey's method for normal means. It is based on applying the Studentized range distribution with $df = \infty$ to a set of approximately standard normally distributed score statistics constructed for the pairs of groups. Schaarschmidt et al. (2008) proposed simultaneous confidence intervals for multiple contrasts of binomial proportions. For discrete small-sample distributions, Tarone (1990) adjusted the Bonferroni method to reduce its conservatism.

Section 3.2: Testing Independence in Two-Way Contingency Tables

3.3 Chi-squared moments/approximations: For hypergeometric sampling for $I \times J$ tables, Haldane (1940) derived $E(X^2)$ and a complex formula for $\text{var}(X^2)$; Dawson (1954) provided a simplified expression. Lewis et al. (1984) derived the third central moment. For 2×2 tables, Pearson (1947) and others since then (e.g., Campbell 2007) suggested using the multiple $(n - 1)/n$ of the chi-squared statistic. For discussion of the adequacy of chi-squared approximations, see Cressie and Read (1989), Read and Cressie (1988) and references therein, Koehler (1986), Koehler and Larntz (1980), Larntz (1978), and Maiste and Weir (2004). Diaconis and Efron (1985) presented inference based on a uniform distribution over all possible tables of the same I , J , and n ; their *volume test* considers the proportion of such tables having $X^2 \leq X^2_0$ for observed value X^2_0 .

3.4 Complex sampling: Social science applications often incorporate clustering and/or stratification. For analyses of categorical data for complex sampling methods and correlated observations, see Bedrick (1983), Cerioli (2002), Fay (1985), Gleser and Moore (1985), Holt et al. (1980), Koch et al. (1975), Koehler and Wilson (1986), LaVange et al. (2001), Rao and Scott (1981, 1987), Rao and Thomas (1988), Scott and Wild (2001), Skinner and Vallet (2010), Tavaré and Altham (1983), and methods of Chapter 13. For example, Gleser and Moore (1985) showed that positive dependence causes null distributions of Pearson statistics to stochastically increase.

3.5 Missing data: Watson (1956) was an early study of effects of missing data in contingency tables. Lipsitz and Fitzmaurice (1996) derived score tests of independence and conditional independence, assuming ignorable nonresponse, and showed that the test statistics have the usual asymptotic chi-squared null distributions. Fleiss et al. (2003, Chap. 16), Little (2005), and Little and Rubin (2002) surveyed ways of dealing with missing data.

3.6 Score and profile likelihood CIs: For other discussion of score test-based intervals, see Agresti (2011) and references therein, Agresti and Ryu (2010), Brown and Li (2005), Gart and Nam (1988), Koopman (1984), Lang (2008), Miettinen and Nurminen (1985), Nurminen (1986), and Exercise 3.27. Cornfield's (1956) interval for the odds ratio utilized a continuity correction. That interval approximates a small-sample interval presented in Section 16.6.4. The Miettinen and Nurminen (1985) score intervals used unbiased variance estimators. For example, their nonnull chi-squared statistic for the difference of proportions has form $[(n - 1)/n][z(\Delta_0)]^2$, so

their interval is slightly wider. Cox and Snell (1989, pp. 51-52) presented the profile likelihood interval for the difference of proportions.

Section 3.4: Two-Way Tables with Ordered Classifications

3.7 Ordinality: Brown and Benedetti (1977) provided null standard errors of ordinal measures appropriate for testing independence. Bhapkar (1968) and Yates (1948) proposed statistics similar to M^2 and statistics for singly ordered tables. Graubard and Korn (1987) listed 14 tests for $2 \times J$ tables that utilize a correlation-type statistic. See also Nair (1987) and Williams (1952). Cohen and Sackrowitz (1992) evaluated decision-theoretic aspects, such as admissibility, of tests based on gamma and local log odds ratios.

Section 3.5: Small-Sample Inference for Contingency Tables

3.8 Continuity correction: For early discussion of Fisher's exact test, see Fisher (1934,1935a,c), Irwin (1935), and Yates (1934). Yates indicated that Fisher suggested the hypergeometric distribution to him for an exact test. He proposed a continuity-corrected version of X^2 for 2×2 tables,

$$X_c^2 = \sum_i \sum_j \frac{(|n_{ij} - \hat{\mu}_{ij}| - 0.5)^2}{\hat{\mu}_{ij}},$$

so that the chi-squared right-tail probability would better approximate the hypergeometric two-sided P -value from Fisher's exact test. Hitchcock (2009) surveyed arguments for and against its use by Yates and other authors. Since software now makes Fisher's exact test feasible even with large samples, this correction is no longer needed.

3.9 Conditional/unconditional: For exact conditional methods, Diaconis and Sturmfels (1998) and Rapallo (2003) proposed algebraic Markov chain algorithms for sampling from the relevant conditional distributions. The controversy over conditioning includes Barnard (1945, 1947, 1949, 1979), Berkson (1978), Cheng et al. (2008), Fisher (1956), Howard (1998), Kempthorne (1979), Little (1989), Lloyd (1988a), Pearson (1947), Rice (1988), Routledge (1992), Suissa and Shuster (1984,1985), and Yates (1934,1984). Discussion of unconditional methods includes Agresti and Min (2001), Berger and Boos (1994), Lin and Yang (2009). Martín Andrés and Silva Mato (1994) summarized and compared various unconditional tests. They found that the method based on the pooled z statistic may not perform as well as Barnard's or Boschloo's test when the sample sizes are very unbalanced. Chan (1998) and Röhmel and Mansmann (1999) considered unconditional tests of equivalence. Zhu and Reid (1994) noted that some information loss about the association occurs in conditioning on the margins except when $\theta = 1$. Other articles on this topic include Berkson (1978), Crook and Good (1980), Gunel and Dickey (1974), Haber (1989), Plackett (1977), and Yates (1984). Agresti (1992,2001) surveyed small-sample methods.

3.10 Two-sided P -value, mid P -value: For discussion of two-sided P -values for Fisher's test, see Blaker (2000), Davis (1986a), Dupont (1986), Lloyd (1988b), Mantel (1987b), and Yates and discussants (1984). For inference using the mid P -value, see Agresti and Gottard (2007) and references therein, Berry and Armitage (1995), Hirji (2005, Sec. 2.5,2.8, and Sec. 2.11.1 for many references), Hwang and Yang (2001), Routledge (1994), Seneta and Phipps (2001), Seneta et al. (1999), Wells (2010), and Yang et al. (2004). Similar benefits can accrue from alternative proposed P -values. One approach, useful when several tables have the same value for a test statistic, uses the table probability to create a more finely partitioned sample space; for tables having the observed test statistic value, only those contribute to the P -value that are no more likely than the observed table (Cohen and Sackrowitz 1992, Kim and Agresti 1995). This depends on more than the sufficient statistic, and in some cases a Rao-Blackwellized version is the mid P -value (Wells 2010). Ordinary P -values obtained with higher-order asymptotic methods without continuity corrections for discreteness yield performance similar to that of the mid P -value (Brazzale et al. 2007, Pierce and Peters 1999, Strawderman and Wells 1998).

Section 3.6: Bayesian Inference for Two-Way Contingency Tables

3.11 Bayes: Agresti and Hitchcock (2005) gave many other references for Bayesian inference in 2×2 tables. Bayes factors for testing independence were considered by Albert (1997), Casella and Moreno (2009), Crook and Good (1980), Good (1976), and Quintana (1998). Altham (1969) used a Bayesian analysis with two ordinal multinomial distributions that evaluates the extent of evidence about stochastic ordering. For situations with no prior information and even an unknown sample space, Walley (1996) proposed an “imprecise Dirichlet model” for multinomial data for which inferences are expressed in terms of posterior upper and lower probabilities that become more precise as the number of observations increases.

EXERCISES

Applications

3.1 A meta-analysis (Moore et al., *Lancet* **370**: 319-328, 2007) of studies on the association between cannabis use (yes, no) and presence of psychosis (yes, no) reported a pooled odds ratio estimate of 1.41, with 95% confidence interval of (1.20, 1.65). Explain how to interpret this interval.

3.2 For 239 golf tournaments on the PGA tour between 2004 and 2009, the economists D. Pope and M. Schweitzer evaluated risk aversion by comparing percentages of putts made when putting for a par versus putting for a birdie (*Am. Econ. Rev.* **101**: 129-157, 2011). For 2828 pairs of putts taken from within 1 inch of each other (from an average distance of about 50 inches) in the same tournament, the sample proportions made were 0.835 when putting for birdie and 0.880 when putting for par (thus avoiding the loss of a bogey). Construct a 95% confidence interval for the difference between the proportions in a corresponding conceptual population. State assumptions, and indicate a key way they do not apply for this study. (Chapter 11 presents more refined methods.)

3.3 [Table 3.10](#) uses the GSS to cross-classify a subject's political party ID with their opinion about whether homosexuals should have the right to marry, for subjects having strong identification with a particular party and strong agreement or disagreement with homosexual marriage. Show that **(a)** $\log(\theta) = 3.728$, **(b)** its standard error is 0.746, and **(c)** the Wald 95% confidence interval for θ is (9.6, 179.3). Name the main factor that causes this interval estimate to be so imprecise.

Table 3.10 Opinion on Homosexual Marriage by Political Party, for Exercise 3.3

Political Party	Homosexuals Should Have Right to Marry	
	Strongly Agree	Strongly Disagree
Strong Democrat	60	44
Strong Republican	2	61

Source: 2010 General Social Survey.

3.4 For [Table 2.10](#) on seat-belt use and results of auto accidents, find and interpret 95% confidence intervals for the conceptual population **(a)** odds ratio, **(b)** difference of proportions, and **(c)** relative risk.

3.5 Refer to [Table 2.5](#) on lung cancer and smoking. Conduct an inferential analysis, and interpret results.

3.6 A study considered the effect of prednisolone on severe hypercalcemia in women with metastatic breast cancer (B. Kristensen et al., *J. Intern. Med.* **232**: 237-245, 1992). Of 30 patients, 15 were randomly selected to receive prednisolone. The other 15 formed a control group. Normalization in their level of serum-ionized calcium was achieved by 7 of the treated patients and none of the control group. Obtain a 95% confidence interval for the odds ratio using **(a)** the Wald interval and **(b)** the profile likelihood. In each case, note the effect of the zero cell count.

3.7 In professional basketball games during 2009-2010, when Kobe Bryant of the Los Angeles Lakers shot a pair of free throws, 8 times he missed both, 152 times he made both, 33 times he made only the first, and 37 times he made only the second. Is it plausible that the successive free throws are independent? (Source of data: www.nba.com and appendix of article 224532 (vol. 6, 2011) by G. Yaari and S. Eisenmann at www.plosone.org investigating the “hot hand” in sports.)

3.8 Refer to Exercise 3.3 and [Table 3.10](#).

- a. Find the z statistic [\(3.12\)](#) and explain how it relates to a chi-squared test.
- b. Find a score or profile likelihood confidence interval for the odds ratio, and compare it to the Wald interval.

3.9 Go to sda.berkeley.edu/GSS and download a contingency table relating attained education and the fundamentalism of one's religious beliefs, for the most recent survey. The GSS variable names are EDUCATION and FUND, and you can enter the year in the Selection Filter, such as YEAR(2010). Using the GSS capabilities or software, conduct the following analyses:

- Report chi-squared statistics, df values, P -values, and interpret.
- Conduct a residual analysis, and interpret. (The standardized residuals are generated by the GSS if you check "Show z statistic" on their menu.)

3.10 As in the previous exercise, download recent GSS data and perform analyses to answer the questions asked.

- Are people happier who believe in life after death? Analyze using the GSS variables HAPPY and POSTLIFE.
- Is belief in the existence of God associated with party ID? Analyze the 3×6 table resulting from using the GSS variables GOD and PARTYID, combining the PARTYID categories 0 and 1 for Democrat, 2, 3, and 4 for Independent, and 5 and 6 for Republican.

3.11 Refer to [Table 3.11](#), GSS data on party ID and race.

Table 3.11 Data for Exercise 3.11 on Party ID and Race

Race	Party Identification		
	Democrat	Independent	Republican
Black	192	75	8
White	459	586	471

Source: 2008 General Social Survey, National Opinion Research Center.

- Using X^2 and G^2 , test the hypothesis of independence between party identification and race. Report the P -values and interpret.
- Use standardized residuals to describe the evidence of association.
- Partition chi-squared into components regarding the choice between Democrat and Independent and between these two combined and Republican. Interpret.

3.12 Using the 2008 GSS, we cross-classified party ID with gender. [Table 3.12](#) shows some results. Explain how to interpret all the results on this printout. (Reschi denotes the Pearson residual and StReschi denotes the standardized residual.)

Table 3.12 Results for Exercise 3.12 on Party ID and Gender

Expected Frequency	dem	indep	repub
female	422	381	273
	393.41	407.05	275.55
male	299	365	232
	327.59	338.95	229.45

Statistic	DF	Value	Prob
Chi-Square	2	8.2943	0.0158
Likelihood Ratio Chi-Square	2	8.3090	0.0157

Observ	Resraw	Reschi	StReschi	Observ	Resraw	Reschi	StReschi
1	28.59	1.58	2.69	4	28.59	1.44	2.69
2	26.05	1.41	2.43	5	-26.05	-1.29	-2.43
3	2.55	0.17	0.26	6	-2.55	-0.15	-0.26

3.13 A recent study (by R. Armenio et al., *J. Am. Dent. Assoc.* **139**: 592–597, 2008) reported results of a double-blind randomized clinical trial comparing tooth sensitivity for 14 patients using a fluoride gel to 15 patients using placebo. Each patient had weekly visits for responses, between 3 and 7 times. The authors reported a 2×2 table having counts (11,57) for placebo and (21,62) for fluoride gel for the (yes, no) response on tooth sensitivity. They reported a P -value of 0.2 for a chi-squared test comparing the two treatments. Discuss the suitability of this analysis. [Hint: Are the observations independent?]

3.14 [Table 3.13](#) classifies a sample of psychiatric patients by their diagnosis and by whether their treatment prescribed drugs. Partition chi-squared into three components to describe differences and similarities among the diagnoses, by comparing (i) the first two rows, (ii) the third and fourth rows, and (iii) the last row to the first and second rows combined and the third

and fourth rows combined.

Table 3.13 Data for Exercise 3.14 on Psychiatric Diagnoses

Diagnosis	Drugs	No Drugs
Schizophrenia	105	8
Affective disorder	12	2
Neurosis	18	19
Personality disorder	47	52
Special symptoms	0	13

Source: Reprinted with permission from E. Helmes and G. C. Fekken, *J. Clin. Psychol.* 42: 569–576, 1986.

3.15 A GSS that cross-classified income in thousands of dollars (<5, 5–15, 15–25, >25) by job satisfaction (very dissatisfied, a little satisfied, moderately satisfied, very satisfied) for black Americans produced a 4×4 table having counts, by row, (2, 4, 13, 3, / 2, 6, 22, 4 / 0, 1, 15, 8 / 0, 3, 13, 8). For this table, $X^2 = 11.5$ ($P = 0.24$), whereas using scores (3, 10, 20, 35) for income and (1, 3, 4, 5) for job satisfaction, $M^2 = 7.04$ ($P = 0.008$). Explain why the results are so different.

3.16 A study on educational aspirations of high school students (S. Crysdale, *Int. J. Compar. Sociol.* 16: 19–36, 1975) measured aspirations with the scale (some high school, high school graduate, some college, college graduate). The student counts in these categories were (9, 44, 13, 10) when family income was low, (11, 52, 23, 22) when family income was middle, and (9, 41, 12, 27) when family income was high.

- a. Test independence of educational aspirations and family income using X^2 or G_2 . Explain the deficiency of this test for these data.
- b. Find the standardized residuals. Do they suggest any association pattern?
- c. Conduct an alternative test that may be more powerful. Interpret.

3.17 Refer to [Table 2.13](#) on homosexual sex and premarital sex.

- a. Construct and interpret a mosaic plot.
- b. Obtain a 95% confidence interval for gamma. Interpret the association.

3.18 [Table 3.14](#) shows the results of a retrospective study comparing radiation therapy with surgery in treating cancer of the larynx. The response indicates whether the cancer was controlled for at least two years following treatment. [Table 3.15](#) shows SAS output. Some R output looks like:

Table 3.14 Data for Exercise 3.18 on Therapy for Cancer of Larynx

	Cancer Controlled	Cancer Not Controlled
Surgery	21	2
Radiation therapy	15	3

Source: Reprinted with permission from W. M. Mendenhall, R. R. Million, D. E. Sharkey, and N. J. Cassisi, *Int. J. Radiat. Oncol. Biol. Phys.* 10: 357–363, 1984, © Pergamon Press.

Table 3.15 SAS Output for Exercise 3.18 on Therapy for Cancer of Larynx

Fisher's Exact Test	
Cell (1, 1) Frequency (F)	21
Left-sided Pr <= F	0.8947
Right-sided Pr >= F	0.3808
Table Probability (P)	0.2755
Two-sided Pr <= P	0.6384

```
< fisher.test(matrix(c(21, 2, 15, 3), ncol=2, byrow=TRUE), alternative="two.sided") p-value = 0.6384
< fisher.test(matrix(c(21, 2, 15, 3), ncol=2, byrow=TRUE), alternative="greater") p-value = 0.3808
< fisher.test(matrix(c(21, 2, 15, 3), ncol=2, byrow=TRUE), alternative="less") p-value = 0.8947
```

- a. Report and interpret the P -value for Fisher's exact test with (i) $H_a: \theta > 1$ and (ii) $H_a: \theta \neq 1$. Explain how the P -values are calculated.
- b. Find and interpret the mid P -value for $H_a: \theta > 1$. Summarize advantages and disadvantages of this type of P -value.

3.19 A study in the Department of Wildlife Ecology at the University of Florida sampled wild common carp fish from a wetland in central Chile. One analysis investigated whether the fish muscle had lead pollutant and whether there was evident malformation in the fish. Of 25 fish without lead, 7 had malformation. Of 14 with lead, 7 had malformation. Report and interpret the P -value for Fisher's exact test for a one-sided alternative of a greater chance of malformation when there is lead pollution.

3.20 Seneta and Phipps (2001) described a medical study that compared subjects with nonacute appendicitis and with acute appendicitis in terms of whether they suffered severe right abdominal pain. Such severe pain was reported by 5 of the 15 nonacute cases and by 1 of the 16 acute cases. The doctors believed that greater density of nerve fibres in the nonacute cases could increase the chance of such pain. Find and interpret the P -value for a one-sided (a) Fisher's exact test and (b) unconditional exact test for two binomials.

3.21 Analyze [Table 3.1](#) using the Bayesian approach with independent uniform prior distributions.

- a. Specify the posterior distribution of (π_1, π_2) .
- b. Using software or your own simulation, estimate the posterior mean of the difference of proportions and find a 95% equal-tail posterior interval. Interpret.

3.22 Refer to the table (11,0 / 0, 1) analyzed with Bayesian methods in Section 3.6.4. Using simulation, estimate $P(\pi_1 > \pi_2 | y_1, n_1; y_2, n_2)$ for independent beta(α_1, α_2) priors having (a) $\alpha_1 = \alpha_2 = 2$, (b) $\alpha_1 = \alpha_2 = 1$, and (c) $\alpha_1 = \alpha_2 = 0.50$. Interpret.

3.23 [Table 3.16](#) cross-classifies votes in the 2000 and 2004 U.S. presidential elections. Treating the two rows as independent binomials and using uniform priors, generate the posterior distribution of the odds ratio. Plot it, and find a 95% equal-tail or HPD posterior interval. What is the disadvantage of an HPD interval here?

Table 3.16 Data on Presidential Votes in 2000 and 2004, for Exercise 3.23

		Vote in 2004	
Political Vote in 2000		Bush	Kerry
Bush		763	65
Gore		59	680

Source: 2006 General Social Survey

Theory and Methods

3.24 Is $\hat{\theta}$ the midpoint of commonly used confidence intervals for the odds ratio θ ? Why or why not?

3.25 For comparing two binomial samples with fixed sample sizes, show that the standard error [\(3.1\)](#) of a log odds ratio increases when, for either sample, the absolute difference of proportions of successes and failures increases. [Hint: Show that the asymptotic variance is minimized when each binomial probability is 0.50. In particular, when an outcome is relatively uncommon, estimates of the log odds ratio tend to be imprecise.]

3.26 Using the delta method as in Section 3.1.6, show that the Wald confidence interval for the logit of a binomial parameter π is

$$\log[\hat{\pi}/(1 - \hat{\pi})] \pm z_{\alpha/2} / \sqrt{n\hat{\pi}(1 - \hat{\pi})}.$$

Explain how to use this interval to obtain one for π itself. [Newcombe (2001) noted that the sample logit is also the midpoint of the score interval [\(1.14\)](#) for π , on the logit scale. He showed that this logit interval contains the score interval.]

3.27 For two parameters, a confidence interval for $\theta_1 - \theta_2$ based on single-sample estimate $\hat{\theta}_i$ and interval (ℓ_i, u_i) for θ_i , $i = 1, 2$, is

$$(\hat{\theta}_1 - \hat{\theta}_2 - \sqrt{(\hat{\theta}_1 - \ell_1)^2 + (u_2 - \hat{\theta}_2)^2}, \quad \hat{\theta}_1 - \hat{\theta}_2 + \sqrt{(u_1 - \hat{\theta}_1)^2 + (\hat{\theta}_2 - \ell_2)^2}).$$

Newcombe (1998b) proposed an interval for $\pi_1 - \pi_2$ using the score interval (ℓ_i, u_i) for π_i that

performs similarly to the score method of Section 3.2.5. It is $(\hat{\pi}_1 - \hat{\pi}_2 - Z_{\alpha/2}s_L, \hat{\pi}_1 - \hat{\pi}_2 + z_{\alpha/2}s_U)$, with

$$s_L = \sqrt{\frac{\ell_1(1-\ell_1)}{n_1} + \frac{u_2(1-u_2)}{n_2}}, \quad s_U = \sqrt{\frac{u_1(1-u_1)}{n_1} + \frac{\ell_2(1-\ell_2)}{n_2}}.$$

Show that this has the general form above of an interval for $\theta_1 - \theta_2$.

3.28 For multinomial sampling, use the asymptotic variance of $\log \hat{\theta}$ to show that for Yule's Q (Exercise 2.38) the asymptotic variance of $\sqrt{n}(\hat{Q} - Q)$ is $(\sum_i \sum_j \pi^{-1}_{ij}) (1 - Q^2)^2/4$ (Yule 1900, 1912).

3.29 For multinomial probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots)$ with a contingency table of arbitrary dimensions, consider a measure of form $g(\boldsymbol{\pi}) = v/\delta$. Show that the asymptotic variance of $\sqrt{n}[g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi})]$ is $\sigma^2 = [\sum_i \pi_i \eta_i^2 - (\sum_i \pi_i \eta_i)^2]/\delta^4$, where $\eta_i = \delta(\partial v/\partial \pi_i) - v(\partial \delta/\partial \pi_i)$ (Goodman and Kruskal 1972).

3.30 Show that $X^2 = n \sum_i \sum_j (p_{ij} - p_{i+}p_{+j})^2 / p_{i+}p_{+j} = n \sum_i \sum_j p_{i+}p_{+j}(a_{ij} - 1)^2$ for the sample association factors $\{a_{ij}\}$. Thus, X^2 can be large when n is large, regardless of whether the association is practically important. Explain why this test, like other tests, merely indicates the degree of evidence against H_0 and does not describe strength of association. ("Like fire, the chi-square test is an excellent servant and a bad master," Sir Austin Bradford Hill, *Proc. R. Soc. Med.* **58**: 295–300, 1965.)

3.31 For a 2×2 table, consider $H_0: \pi_{11} = \theta_2, \pi_{12} = \pi_{21} = \theta(1 - \theta), \pi_{22} = (1 - \theta)^2$.

- a. Show that the marginal distributions are identical and that independence holds.
- b. For a multinomial sample, under H_0 show that $\hat{\theta} = (p_{1+} + p_{+1})/2$.
- c. Explain how to test H_0 . Show that $df = 2$ for the test statistic.
- d. Refer to Exercise 3.7. Are Kobe Bryant's pairs of free throws plausibly independent *and* identically distributed?

3.32 For testing independence, show that $X^2 \leq n \min(I-1, J-1)$. Hence $V^2 = X^2/[n \min(I-1, J-1)]$ falls between 0 and 1 (Cramér 1946). [For 2×2 tables, X^2/n is often called *phi-squared*; it equals Goodman and Kruskal's tau of Exercise 2.39. Other measures based on X^2 include the *contingency coefficient* $[X^2 / (X^2 + n)]^{1/2}$, which Pearson (1904) proposed as an estimate of the correlation for an underlying bivariate normal distribution.]

3.33 For a 1×2 table (i.e., a single binomial Y based on n trials, with probabilities π and $1 - \pi$), consider testing $H_0: \pi = \pi_0$.

- a. Show that the Pearson residuals are

$$(y - n\pi_0)/\sqrt{n\pi_0} \quad \text{and} \quad -(y - n\pi_0)/\sqrt{n(1 - \pi_0)},$$

which have differing absolute values when $\pi_0 \neq 0.50$.

- b. Show that the standardized residuals are

$$(y - n\pi_0)/\sqrt{n\pi_0(1 - \pi_0)} \quad \text{and} \quad -(y - n\pi_0)/\sqrt{n\pi_0(1 - \pi_0)}.$$

Explain why these are more suitable than Pearson residuals.

3.34 For a 2×2 table, show that:

- a. The four Pearson residuals may take different values.
- b. All four standardized residuals have the same absolute value. (This is sensible, since $df = 1$.)
- c. The square of each standardized residual equals X^2 .

3.35 Use a partitioning argument to explain why G^2 for testing independence cannot increase after combining two rows (or two columns) of a contingency table. [Hint: Explain why G^2 for full table = G^2 for collapsed table + G^2 for table of the two rows that are combined in the collapsed table.]

3.36 Assume independence, and let $p_{ij} = n_{ij}/n$ and $\hat{\pi}_{ij} = p_{i+}p_{+j}$.

a. Show that p_{ij} and $\hat{\pi}_{ij}$ are unbiased for $\pi_{ij} = \pi_{i+}\pi_{+j}$.

b. Show that $\text{var}(p_{ij}) = \pi_{i+}\pi_{+j}(1 - \pi_{i+}\pi_{+j})/n$.

c. Using $E(p_{i+}P_{+j})^2 = E(p_{i+}^2)E(P_{+j}^2)$ and $E(p_{i+}^2) = \text{var}(p_{i+}) + [E(p_{i+})]^2$, show that

$$\begin{aligned}\text{var}(\hat{\pi}_{ij}) &= \{\pi_{i+}\pi_{+j}[\pi_{i+}(1 - \pi_{+j}) + \pi_{+j}(1 - \pi_{i+})]\}/n \\ &\quad + \pi_{i+}(1 - \pi_{i+})\pi_{+j}(1 - \pi_{+j})/n^2.\end{aligned}$$

d. As $n \rightarrow \infty$, show that $\lim \text{var}(\sqrt{n}\hat{\pi}_{ij}) \leq \lim \text{var}(\sqrt{n}p_{ij})$, with equality only if $\pi_{ij} = 1$ or 0.

Hence, if the model holds or if it nearly holds, the model estimator is better than the sample proportion.

3.37 Consider an $I \times J$ table with ordered columns and unordered rows. *Ridits* (Bross 1958) are data-based column scores. The j th sample ridit is the average cumulative proportion within category j ,

$$\hat{r}_j = \sum_{k=1}^{j-1} p_{+k} + \left(\frac{1}{2}\right) p_{+j}.$$

The sample mean ridit in row i is $\hat{R}_i = \sum_j \hat{r}_j p_{j|i}$. Show that $\sum_j p_{+j}\hat{r}_j = 0.50$ and $\sum_i p_{i+}\hat{R}_i = 0.50$. [For ridit analyses, see Agresti (2010, Sec. 2.1), Beder and Heim (1990), Bross (1958), Fleiss et al. (2003, Sec. 9.4), and Landis et al. (1978).]

3.38 Show that the sample value of the uncertainty coefficient (2.13) satisfies $\vartheta = -G^2/2n$ ($\sum p_{+j} \log p_{+j}$). [Haberman (1982) gave its standard error.]

3.39 Of six candidates for three managerial positions, denote the females by F1, F2, F3 and the males by M1, M2, M3.

a. Identify the 20 possible combinations of candidates that could be selected. Construct the contingency table for the actual sample, which is (F2, M1, M3).

b. Let p_1 denote the sample proportion of males selected and p_2 the sample proportion of females selected. Of the 20 possible samples, show that 10 have $p_1 - p_2 \geq \frac{1}{3}$. Thus, if the three managers were randomly selected, $P(p_1 - P_2 \geq \frac{1}{3}) = 10/20 = 0.50$. Explain why this is the P -value for Fisher's exact test with $H_a: \pi_1 > \pi_2$.

3.40 When a test statistic has a continuous distribution, the P -value has a null uniform distribution, $P(P\text{-value} \leq \alpha) = \alpha$ for $0 < \alpha < 1$. For Fisher's exact test, explain why $P(P\text{-value} \leq \alpha) \leq \alpha$. [Hint: $P(P\text{-value} \leq \alpha) = E[P(P\text{-value} \leq \alpha | n_{1+}, n_{+1}, n)]$.]

3.41 Note 3.3 showed moments of the hypergeometric distribution (3.17). Letting $\rho = n_{+1}/n$, show that n_{11} has the same mean as a binomial random variable for n_{1+} trials with success probability ρ , and that it has its variance multiplied by a finite population correction factor $(n - n_{1+})/(n - 1)$. (The hypergeometric is similar to the binomial when n_{1+} is small compared to n .)

3.42 For the tea-tasting data (Table 3.9), construct the null distributions of the ordinary P -value and the mid P -value for Fisher's exact test with $H_a: \theta > 1$. Find and compare their expected values.

3.43 In Section 3.5.6 we analyzed a 2×2 table having entries (3, 0 / 0, 3). Explain why the unconditional P -value, evaluated at $\pi = 0.50$, is related to Fisher conditional P -values for various tables by

$$P(X^2 \geq 6) = \sum_{k=0}^6 P(X^2 \geq 6 | n_{+1} = k) P(n_{+1} = k).$$

Thus, the unconditional P -value of $\frac{1}{32}$ is a weighted average of the Fisher P -value for the observed column margins and P -values of 0 corresponding to the impossibility of getting results as extreme as observed if other margins had occurred (i.e., $\frac{1}{32} = 0.10 \left[\binom{6}{3} \left(\frac{1}{2}\right)^6 \right]$). The Fisher quote in Section 3.5.7 gave his view about this.

3.44 For testing $H_0: \pi_1 = \pi_2$ with two binomial variates y_1 and y_2 , a “reasonable” test would not

reject H_0 if $y_1 = y_2 = 0$. Since as π_1 and π_2 approach 0, the probability of this converges to 1 even if $\pi_1 \neq \pi_2$, explain why any such test is biased, potentially having power less than its size (Haber 1986).

3.45 For independent uniform prior distributions for two binomial parameters, show that $r = \pi_1/\pi_2$ has prior density $g(r) = \frac{1}{2}$ for $0 \leq r \leq 1$ and $g(r) = 1/2r^2$ for $r > 1$.

3.46 Explain why a Bayesian HPD interval is sensible for $\pi_1 - \pi_2$ but not usually for π_1/π_2 .

3.47 Consider a particular choice of Dirichlet means $\{\gamma_{ij} = E(\pi_{ij}) = \alpha_{ij}/K\}$ for the Bayes estimator (1.19) extended to two-way tables. Show that the total mean squared error is

$$[K/(n+K)]^2 \left[\sum (\pi_{ij} - \gamma_{ij})^2 \right] + [n/(n+K)]^2 \left[1 - \sum \pi_{ij}^2 \right].$$

Show that the value of K that minimizes this is

$$K = \left(1 - \sum \pi_{ij}^2 \right) / \left[\sum (\gamma_{ij} - \pi_{ij})^2 \right].$$

Fienberg and Holland (1973) showed this and used the empirical Bayes approach of estimating K by replacing $\boldsymbol{\pi}$ by the sample proportion \mathbf{p} and letting $\{\gamma_{ij} = p_{i+}p_{+j}\}$. Albert (2010) surveyed Bayesian methods for smoothing contingency tables.

¹This is also true for ML estimators of model parameters presented in later chapters.

²See www.stat.ufl.edu/~aa/cda/cda.html.

³In Section 10.2.4 we explain why this partitioning works.

⁴For example, PROC FREQ in SAS.

⁵In Section 5.3.8 we present the theory behind such a power comparison.

⁶See Note 5.7 for efficiency results when one variable is binary.

⁷See www.stat.ufl.edu/~aa/cda/R/bayes/index.html for R functions to construct posterior intervals for the odds ratio, relative risk, and difference of proportions.

CHAPTER 4

Introduction to Generalized Linear Models

In Chapters 2 and 3 we focused on methods for two-way contingency tables. Most studies, however, have several explanatory variables, and they may be continuous as well as categorical. *Modeling* helps us to efficiently evaluate effects and provide improved estimates of response probabilities because of the parsimonious reduction in the number of parameters.

The rest of the book focuses on model building for categorical response variables. In this chapter we introduce a family of *generalized linear models* that contains important models for categorical responses as well as standard models for continuous responses. Section 4.1 defines three components common to all generalized linear models. Section 4.2 illustrates with models for binary responses. The most important case is *logistic regression*, a linear model for the log odds (*logit*) transformation of a binomial parameter. In Chapters 5 through 8 we study these models in detail.

In Section 4.3 we present generalized linear models for counts. A *Poisson regression model* called a *loglinear model* is a linear model for the log of a Poisson mean. In Chapters 9 and 10 we use them for modeling counts in contingency tables having multiple response variables.

Sections 4.4 through 4.7 are more technical. Readers wanting mainly an overview of methods can skip them or read them lightly. Section 4.4 shows likelihood equations and the asymptotic covariance matrix of maximum likelihood (ML) parameter estimates for generalized linear models, and Section 4.5 summarizes inferential methods. Methods of solving the likelihood equations are presented in Section 4.6. In the final section we introduce a generalization, *quasi-likelihood*, that further extends the scope of models.

4.1 THE GENERALIZED LINEAR MODEL

Generalized linear models (GLMs) extend ordinary regression models to encompass nonnormal response distributions and modeling functions of the mean. They have three components: A *random component* identifies the response variable Y and its probability distribution; a *systematic component* specifies explanatory variables used in a linear predictor function; and a *link function* specifies the function of $E(Y)$ that the model equates to the linear predictor. Nelder and Wedderburn (1972) introduced the class of GLMs, although the most important models in the class were established before then.

4.1.1 Components of Generalized Linear Models

The *random component* of a GLM consists of a response variable Y with independent observations (y_1, \dots, y_N) from a distribution in the natural exponential family. This family has probability density function or mass function of form

$$(4.1) \quad f(y_i; \theta_i) = a(\theta_i)b(y_i)\exp[y_i Q(\theta_i)].$$

Several important distributions are special cases, including the Poisson and binomial. The value of the parameter θ_i , varies for $i = 1, \dots, N$ as a function of values of explanatory variables. The parameter $Q(\theta)$ is called the *natural parameter*. In Section 4.4 we present a more general formula (4.17) for f that also permits a dispersion parameter, but (4.1) is sufficient for the discrete data models that are the primary focus of this book.

The *systematic component* of a GLM relates a vector (η_1, \dots, η_N) to the explanatory variables through a linear model. Let x_{ij} denote the value of explanatory variable j ($j = 0, 1, 2, \dots$) for subject i . Then

$$\eta_i = \sum_j \beta_j x_{ij}, \quad i = 1, \dots, N.$$

This linear combination of explanatory variables is called the *linear predictor*. Usually, $x_{i0} = 1$ for all i , representing the coefficient of an intercept term β_0 (often denoted by α) in the model.

The third component of a GLM is a *link function* that connects the random and systematic components. Let $\mu_i = E(Y_i)$, $i = 1, \dots, N$. The model links μ_i to η_i by $\eta_i = g(\mu_i)$, where the link function g is a monotonic, differentiable function. Thus, g links μ_i to explanatory variables through the formula

$$(4.2) \quad g(\mu_i) = \sum_j \beta_j x_{ij}, \quad i = 1, \dots, N.$$

The link function $g(\mu) = \mu$, called the *identity link*, has $\eta_i = \mu_i$. It specifies a linear model for the mean itself. This is the link function for ordinary regression with normally distributed Y . The link function that transforms the mean to the natural parameter is called the *canonical link*. For it, $g(\mu_i = Q(\theta_i))$, and $Q(\theta_i) = \sum_j \beta_j x_{ij}$. Sections 4.1.2 and 4.1.3 show examples.

In summary, a GLM is a linear model for a transformed mean of a response variable that has distribution in the natural exponential family. We now illustrate the three components by introducing the key GLMs for discrete response variables.

4.1.2 Binomial Logit Models for Binary Data

Many response variables are binary. We represent the “success” and “failure” outcomes by 1 and 0. A *Bernoulli trial* has probabilities $P(Y=1) = \pi$ and $P(Y=0) = 1 - \pi$, for which $E(Y) = \pi$. This is the special case of the binomial distribution (1.1) with $n = 1$. We can express the probability mass function as

$$\begin{aligned} f(y; \pi) &= \pi^y (1 - \pi)^{1-y} = (1 - \pi)[\pi/(1 - \pi)]^y \\ (4.3) \quad &= (1 - \pi) \exp \left[y \left(\log \frac{\pi}{1 - \pi} \right) \right] \end{aligned}$$

for $y = 0$ and 1. This is in the natural exponential family (4.1), identifying θ with π , $a(\pi) = 1 - \pi$, $b(y) = 1$, and $Q(\pi) = \log[\pi/(1 - \pi)]$. The natural parameter $\log[\pi/(1 - \pi)]$ is the log odds of response outcome 1, the *logit* of π . This is the canonical link function. GLMs using the logit link are introduced further in Section 4.2.3. They are referred to as *logistic regression models*, or sometimes simply as *logit models*.

4.1.3 Poisson Loglinear Models for Count Data

Some response variables have counts as their possible outcomes. In a health survey, each observation might be the number of illnesses in the past year for which the subject visited a doctor. Counts also occur as entries in contingency tables.

The simplest distribution for count data is the Poisson. The Poisson probability mass function (1.4) for a count Y is

$$f(y; \mu) = \frac{e^{-\mu} \mu^y}{y!} = \exp(-\mu) \left(\frac{1}{y!} \right) \exp[y(\log \mu)], \quad y = 0, 1, 2, \dots$$

This has natural exponential form (4.1) with $\theta = \mu$, $a(\mu) = \exp(-\mu)$, $b(y) = 1/y!$, and $Q(\mu) = \log \mu$. The natural parameter is $\log \mu$, so the canonical link function is the log link, $\eta = \log \mu$. The model using this link function is

$$(4.4) \quad \log \mu_i = \sum_j \beta_j x_{ij}, \quad i = 1, \dots, N.$$

This model, to be introduced further in Section 4.3.1, is called a *Poisson loglinear model*.

4.1.4 Generalized Linear Models for Continuous Responses

The class of GLMs also includes models for continuous responses. The normal distribution is in a natural exponential family that includes dispersion parameters. Its natural parameter is the mean. Therefore, an ordinary regression model is a GLM using the identity link. [Table 4.1](#) lists this and other standard models for a normal random component. The table also lists GLMs for discrete responses that are presented in Chapters 5–10.

[Table 4.1](#) Types of Generalized Linear Models for Statistical Analysis

Random Component	Link Function	Systematic Component	Model	Chapters
Normal	Identity	Continuous	Regression	
Normal	Identity	Categorical	Analysis of variance	
Normal	Identity	Mixed	Analysis of covariance	
Binomial	Logit	Mixed	Logistic regression	5 and 6
Binomial	Probit and others	Mixed	Binary regression	7
Multinomial	Generalized logit	Mixed	Multinomial response	8
Poisson	Log	Mixed	Loglinear	9 and 10

4.1.5 Deviance of a GLM

For a particular GLM with observations $\mathbf{y} = (y_1, \dots, y_N)$, let $L(\boldsymbol{\mu}; \mathbf{y})$ denote the log-likelihood function expressed in terms of the means $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$. Let $L(\hat{\boldsymbol{\mu}}; \mathbf{y})$ denote the maximum of the log likelihood for the model. Considered for all possible models, the maximum achievable log likelihood is $L(\mathbf{y}; \mathbf{y})$. This occurs for the most general model, having a separate parameter for each observation and the perfect fit $\hat{\boldsymbol{\mu}} = \mathbf{y}$. Such a model is called the *saturated model*. This model is not useful, because it does not provide data reduction. However, it serves as a baseline for comparison with other model fits.

The *deviance* of a Poisson or binomial GLM is defined to be

$$-2[L(\hat{\boldsymbol{\mu}}; \mathbf{y}) - L(\mathbf{y}; \mathbf{y})].$$

This is the likelihood-ratio statistic for testing the null hypothesis that the model holds against the general alternative (i.e., the saturated model). We use the deviance throughout the book for model checking and for inferential comparisons of models. Methods for analyzing the deviance generalize analysis of variance methods for normal linear models.

For some applications with Poisson and binomial GLMs, the number of observations N is fixed and the individual counts are relatively large. Then the deviance has an approximate chi-squared null distribution. The $df = N - p$, where p is the number of model parameters; that is, df equals the difference between the numbers of parameters in the saturated model and in the unsaturated model. The deviance then provides a test of model fit.

One such example is independent binomial counts at N fixed settings of predictors when the number of trials at each setting is large. Let Y_i be $\text{bin}(n_i, \pi_i)$, $i = 1, \dots, N$. Consider the simple model of homogeneity, $\pi_i = \alpha$ all i . It has $p = 1$ parameter. The saturated model makes no assumption about $\{\pi_i\}$, letting them be any N values between 0 and 1.0. It has N parameters. The deviance for the homogeneity model has $df = N - 1$. In fact, it equals the G^2 likelihood-ratio statistic (3.11) for testing independence in the $N \times 2$ contingency table that these samples form. Under independence, its distribution converges to a chi-squared distribution as the $\{n_i\}$ increase, for fixed N . Another example is a contingency table constructed from sample survey data, in which the classification categories and the number of cells N is fixed as we collect more data, and we treat the cell counts as Poisson variates.

4.1.6 Advantages of GLMs Versus Transforming the Data

A traditional way to model data transforms Y so that it has approximately a normal distribution with constant variance; then, ordinary least-squares regression is applicable. With GLMs, by contrast, the choice of link function is separate from the choice of random component. If a link is useful in the sense that a linear model for the predictors is plausible for that link, it is not necessary that it also stabilizes variance or produces normality. This is because the fitting process maximizes the likelihood for the choice of distribution for Y , and that choice is not restricted to normality.

Let g denote a function, such as the log function, that is a link function in the GLM approach or a transformation function in the transformed data approach. An advantage of the GLM formulation is that the model parameters describe $g[E(Y)]$, rather than $E[g(Y)]$ as in the transformed data approach. With the GLM approach, those parameters also describe effects of explanatory variables on $E(Y)$, after applying the inverse function for g . Such effects are more relevant than effects of explanatory variables on $E[g(Y)]$.

GLMs provide a unified theory of modeling that encompasses the most important models for continuous and discrete variables. Models studied in this text are GLMs with binomial or Poisson random component, or multivariate extensions of GLMs. The ML parameter estimates are computed with an algorithm, presented in Section 4.6, that iteratively uses a weighted version of least squares. A reason for restricting GLMs to the exponential family of distributions for Y is that the same algorithm applies to this entire family, for any choice of link function.

Nearly all statistical software has the facility to fit GLMs. This text's computing appendix at www.stat.ufl.edu/~aa/cda/cda.html gives details.

4.2 GENERALIZED LINEAR MODELS FOR BINARY DATA

Let Y denote a binary response variable, such as the result of a medical treatment (success, failure). Each observation has one of two outcomes, denoted by 1 and 0, which we treat as a binomial variate for a single Bernoulli trial. The mean $E(Y) = P(Y = 1)$. We denote $P(Y = 1)$ by $\pi(\mathbf{x})$, reflecting its dependence on values $\mathbf{x} = (x_1, \dots, x_p)$ of explanatory variables. The variance of Y is

$$\text{var}(Y) = \pi(\mathbf{x})[1 - \pi(\mathbf{x})],$$

which is the binomial variance for $n = 1$.

4.2.1 Linear Probability Model

For a binary response variable, the regression model

$$(4.5) \quad \pi(\mathbf{x}) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p$$

is called a *linear probability model*. With independent observations it is a GLM with binomial random component and identity link function.

This model has a major structural defect: Probabilities fall between 0 and 1, but linear functions take values over the entire real line. Model (4.5) can have $\pi(\mathbf{x}) < 0$ and/or $\pi(\mathbf{x}) > 1$ for some \mathbf{x} values. The model can be valid over a restricted range of \mathbf{x} values. When it is plausible, an advantage is its simple interpretation: β_j is the change in $\pi(\mathbf{x})$ for a one-unit increase in x_j .

We defer to Section 4.6 the technical details of ML model fitting for this and other GLMs. Since $\text{var}(Y) = \pi(\mathbf{x})[1 - \pi(\mathbf{x})]$, the variance depends on \mathbf{x} through its influence on $\pi(\mathbf{x})$. The constant variance condition that makes ordinary least-squares estimators optimal (i.e., minimum variance in the class of linear unbiased estimators) is not satisfied, so the ML estimator is more efficient than least squares. The estimates and standard errors for ML and least squares are usually similar, however, when $\hat{\pi}(\mathbf{x})$ for the sample \mathbf{x} values falls in the range within which the variance is relatively stable, about 0.3 to 0.7. When used with multiple explanatory variables, difficulties often occur with ML model fitting because at a step of the iterative fitting process, $\hat{\pi}(\mathbf{x})$ falls outside the [0, 1] range for some subjects' \mathbf{x} values. Least-squares fitting still works in such cases, but also typically gives such unsatisfactory $\hat{\pi}(\mathbf{x})$ estimates. Also Y , being binary, is very far from normally distributed, so the usual t sampling distribution for standardized least-squares estimators do not apply.

4.2.2 Example: Snoring and Heart Disease

We illustrate the linear probability model with [Table 4.2](#), from an epidemiological survey to investigate snoring as a risk factor for heart disease. The sample consists of 2484 subjects who visited four family practice units in Toronto that served different socioeconomic classes and ethnic groups. Those surveyed were classified according to their spouses' report of how much they snored and according to whether they reported having heart disease. The model states that the probability of heart disease is linearly related to the level of snoring x . We treat the rows of the table as independent binomial samples. No obvious choice of scores exists for categories of x . We used (0, 2, 4, 5), treating the last two levels as closer than the other adjacent pairs. ML estimates and standard errors are the same if we use a data file of 2484 binary observations or if we enter the four binomial totals of "yes" and "no" responses listed in [Table 4.2](#).

Table 4.2 Relationship Between Snoring and Heart Disease

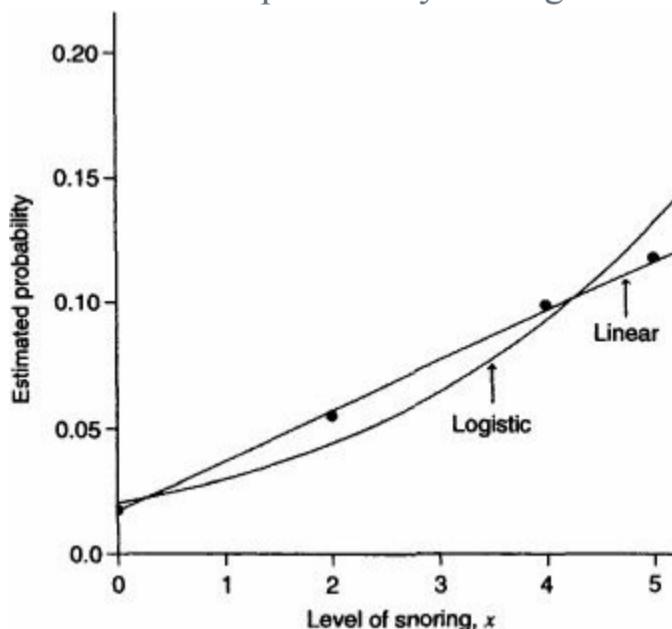
Snoring	Heart Disease		Proportion Yes	Linear Fit ^a	Logistic Fit ^a
	Yes	No			
Never	24	1355	0.017	0.017	0.021
Occasionally	35	603	0.055	0.057	0.044
Nearly every night	21	192	0.099	0.096	0.093
Every night	30	224	0.118	0.116	0.132

^aModel fits refer to proportion of yes responses.

Source: P. G. Norton and E. V. Dunn, *Br. Med. J.* 291: 630–632, 1985, BMJ Publishing Group.

Software reports the ML fit, $\hat{\pi}(x) = 0.0172 + 0.0198x$, with $\hat{\beta} = 0.0198$ having $SE = 0.0028$. For nonsnorers ($x = 0$), the estimated proportion of subjects having heart disease is 0.0172. We refer to the estimated values of $E(Y)$ for a GLM as *fitted values*. [Table 4.2](#) shows the sample proportions and the fitted values for this model. [Figure 4.1](#) graphs the sample and fitted values. The table and graph suggest that the model fits well. (In Section 5.2.3 we present formal goodness-of-fit analyses for binary-response GLMs.) The model interpretation is simple. The estimated probability of heart disease is about 0.02 for nonsnorers; it increases $2(0.0198) = 0.04$ for occasional snorers, another 0.04 for those who snore nearly every night, and another 0.02 for those who always snore. The study was observational, and it is unclear whether this association could be due to some confounding factor or a medical condition such as sleep apnea.

Figure 4.1 Estimated probabilities for linear probability and logistic regression models.

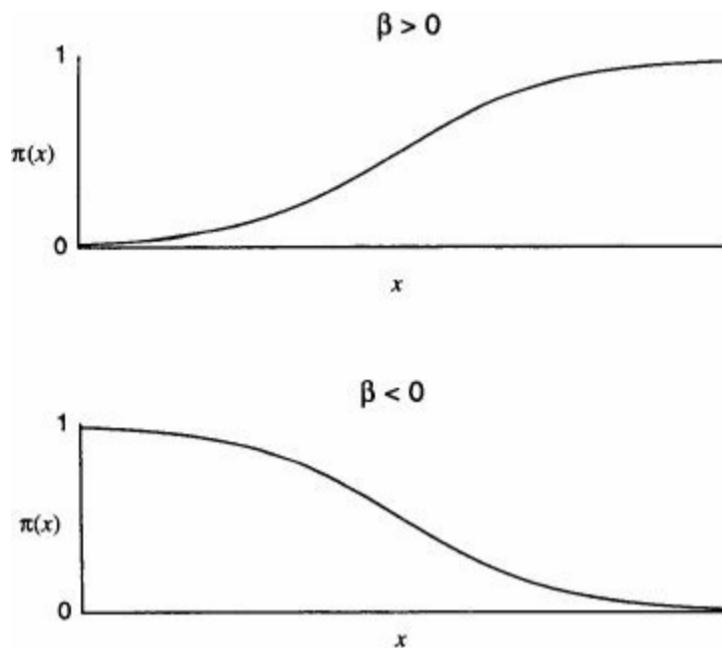


4.2.3 Logistic Regression Model

Usually, binary data result from a *nonlinear* relationship between $\pi(x)$ and x . A fixed change in x often has less impact when $\pi(x)$ is near 0 or 1 than when $\pi(x)$ is near 0.50. In the purchase of an automobile, consider the choice between buying new or used. Let $\pi(x)$ denote the probability of selecting new when annual family income = x . An increase of \$10,000 in annual income would have less effect when $x = \$1,000,000$ [for which $\pi(x)$ is near 1] than when $x = \$50,000$.

In practice, nonlinear relationships between $\pi(x)$ and x are often monotonic, with $\pi(x)$ increasing continuously or $\pi(x)$ decreasing continuously as x increases. The S-shaped curves in [Figure 4.2](#) are typical. The most important curve with this shape has the model formula

[Figure 4.2](#) Logistic regression functions.



$$(4.6) \quad \pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}.$$

This is a *logistic regression* model. As x increases, $\pi(x)$ increases when $\beta > 0$ and decreases when $\beta < 0$.

Let's find the link function for which logistic regression is a GLM. For [\(4.6\)](#) extended to multiple predictors, the odds are

$$\frac{\pi(x)}{1 - \pi(x)} = \exp(\alpha + \beta_1 x_1 + \cdots + \beta_p x_p).$$

The log odds has the linear relationship

$$(4.7) \quad \log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Thus, the appropriate link is the log odds transformation, the *logit*. Logistic regression models are GLMs with binomial random component and logit link function.

The logit is the natural parameter for the binomial distribution, so the logit link is its canonical link function. Whereas $\pi(x)$ must fall in the (0, 1) range, the logit can be any real number. The real numbers are also the range for linear predictors that form the systematic component of a GLM. So, this model does not have the structural problem that the linear probability model has.

For the snoring data in [Table 4.2](#), software reports the logistic regression ML fit

$$\text{logit}[\hat{\pi}(x)] = -3.87 + 0.40x.$$

The positive $\hat{\beta} = 0.40$ reflects the increased incidence of heart disease at higher snoring levels. In Chapters 5 and 6 we study logistic regression in detail and interpret such equations. Estimated probabilities result from substituting x values into the estimate of probability formula [\(4.6\)](#). [Table 4.2](#) also reports these fitted values. [Figure 4.1](#) displays the fit. The fit is close to linear over this narrow range of estimated probabilities, and results are similar to those for the linear probability model.

4.2.4 Binomial GLM for 2×2 Contingency Tables

Among the simplest GLMs for a binary response is the one having a single explanatory variable x that is also binary. Label its values by 0 and 1. For a given link function, the GLM

$$\text{link}[\pi(x)] = \alpha + \beta x$$

has the effect of x described by

$$\beta = \text{link}[\pi(1)] - \text{link}[\pi(0)].$$

For the identity link, $\beta = \pi(1) - \pi(0)$ is the difference between proportions. For the log link, $\beta = \log[\pi(1)] - \log[\pi(0)] = \log[\pi(1)/\pi(0)]$ is the log relative risk. For the logit link,

$$\begin{aligned}\beta &= \text{logit}[\pi(1)] - \text{logit}[\pi(0)] = \log \frac{\pi(1)}{1 - \pi(1)} - \log \frac{\pi(0)}{1 - \pi(0)} \\ &= \log \left[\frac{\pi(1)/(1 - \pi(1))}{\pi(0)/(1 - \pi(0))} \right]\end{aligned}$$

is the log odds ratio. Measures of association for 2×2 tables are effect parameters in GLMs for binary data.

4.2.5 Probit and Inverse cdf Link Functions

A monotone regression curve such as the first one in [Figure 4.2](#) has the shape of a cumulative distribution function (cdf) for a continuous random variable. This suggests a model for a binary response having form $\pi(x) = F(x)$ for some cdf F .

Using a class of location-scale cdf's, such as normal cdf's with their variety of means and variances, permits the curve $\pi(x) = F(x)$ to have flexibility in the rate of increase and in the location where most of that increase occurs. Let $\Phi(\cdot)$ denote the standard cdf of the class, such as the $N(0, 1)$ cdf. Using Φ but writing the model as

$$(4.8) \quad \pi(x) = \Phi(\alpha + \beta x)$$

provides the same flexibility. The values of α and β determine the particular cdf in the class. Replacing x by βx permits the curve to increase at a different rate than the standard cdf (or even to decrease if $\beta < 0$); varying α moves the curve to the left or right.

When Φ is strictly increasing over the entire real line, its inverse function Φ^{-1} exists and [\(4.8\)](#) is, equivalently,

$$(4.9) \quad \Phi^{-1}[\pi(x)] = \alpha + \beta x.$$

For this class of cdf shapes, the link function for the GLM is Φ^{-1} . The link function maps the $(0, 1)$ range of probabilities onto $(-\infty, \infty)$, the range of linear predictors. The curve has the shape of a normal cdf when Φ is the standard normal cdf. Model [\(4.9\)](#) is then called the *probit* model. This curve has similar appearance to the logistic regression curve. Probit models are discussed in Section 7.1.

When $\beta > 0$, the logistic regression curve [\(4.6\)](#) is a cdf for the *logistic distribution*. When $\beta < 0$, the curve for $1 - \pi(x)$ has that appearance. The cdf of the logistic distribution with mean μ and dispersion parameter $\tau > 0$ is

$$F(x) = \frac{\exp[(x - \mu)/\tau]}{1 + \exp[(x - \mu)/\tau]}, \quad -\infty < x < \infty.$$

The corresponding probability density function (pdf) is symmetric and bell-shaped, with standard deviation $\tau\pi/\sqrt{3}$, for the mathematical constant $\pi = 3.14 \dots$. It looks much like the normal density with the same mean and standard deviation but with slightly thicker tails.¹ The standard form of the logistic cdf has $\mu = 0$ and $\tau = 1$, so $\Phi(x) = e^x/(1 + e^x)$. For that function, the logistic regression curve [\(4.6\)](#) has form $\pi(x) = \Phi(\alpha + \beta x)$. By [\(4.9\)](#) the logit transformation is simply the inverse function for the standard logistic cdf; that is, when $\Phi(x) = \pi(x) = e^x/(1 + e^x)$, then $x = \Phi^{-1}[\pi(x)] = \log[\pi(x)/(1 - \pi(x))]$.

4.2.6 Latent Tolerance Motivation for Binary Response Models

We now present another motivation for the link function having the form of the inverse of a cdf. It results from early applications of binary response models to toxicology studies, such as in Bliss (1935), with an unobserved *tolerance distribution*.

In toxicology, binary response models describe the effect of dosage of a toxin on whether a subject dies. Let x denote the dosage level. For a randomly selected subject, let $Y = 1$ if the subject dies. Suppose that the subject has a tolerance threshold T for the dosage, with ($Y = 1$) equivalent to ($T \leq x$). For instance, an insect survives if the dosage x is less than T and dies if the dosage is at least T . Tolerances vary among subjects, and let $F(t) = P(T \leq t)$. For fixed dosage x , the probability a randomly selected subject dies is

$$\pi(x) = P(Y = 1|X = x) = P(T \leq x) = F(x).$$

That is, the appropriate binary model is the one having the shape of the cdf F of the tolerance distribution.

An unobserved variable such as T is referred to as a *latent variable*. In practice we do not know the particular F that generates T , and we assume that F belongs to some parametric family. Let Φ denote the standard cdf for that family. A common standardization uses the mean and standard deviation of T , so that

$$\pi(x) = F(x) = \Phi[(x - \mu)/\sigma].$$

Then, the model has form $\pi(x) = \Phi(\alpha + \beta x)$, for $\alpha = -\mu/\sigma$ and $\beta = 1/\sigma$. In GLM form,

$$(4.10) \quad \Phi^{-1}[\pi(x)] = \alpha + \beta x.$$

Whereas the cdf maps the real line onto the (0, 1) probability scale, the inverse cdf maps the (0, 1) scale for $\pi(x)$ onto the real line values for linear predictors in binary response models.

4.3 GENERALIZED LINEAR MODELS FOR COUNTS AND RATES

The best known GLMs for count data assume a Poisson distribution for Y . We'll use Poisson GLMs for counts in contingency tables with categorical response variables. We first introduce Poisson GLMs to model count or rate data for a single nonnegative integer-valued response variable.

4.3.1 Poisson Loglinear Models

The Poisson distribution (1.4) has a positive mean μ . Although a GLM can model a positive mean using the identity link, it is more common to model the log of the mean. Like the linear predictor, the log mean can take any real value. The log mean is the natural parameter for the Poisson distribution, and the log link is the canonical link for a Poisson GLM. A Poisson loglinear GLM assumes a Poisson distribution for Y and uses the log link.

The Poisson loglinear model with explanatory variables \mathbf{x} is

$$(4.11) \log \mu(\mathbf{x}) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p.$$

For this model, the mean satisfies the exponential relationship

$$(4.12) \mu(\mathbf{x}) = \exp(\alpha + \beta_1 x_1 + \cdots + \beta_p x_p) = e^\alpha (e^{\beta_1})^{x_1} \cdots (e^{\beta_p})^{x_p}.$$

A 1-unit increase in x_j has a multiplicative impact of e^{β_j} : The mean at $x_j + 1$ equals the mean at x_j multiplied by e^{β_j} .

4.3.2 Example: Horseshoe Crab Mating

We illustrate Poisson GLMs using a study of female horseshoe crabs² on an island in the Gulf of Mexico. During spawning season, the females migrate to a shore to breed, with a male attached to her posterior spine, and she burrows into the sand and lays clusters of eggs. During spawning, other male crabs may group around the pair and may also fertilize the eggs. These male crabs that cluster around the female crab are called *satellites*.

In this example, the response outcome for each of 173 female crabs is her number of satellites. Explanatory variables are the female crab's color, spine condition, weight, and carapace width. [Table 4.3](#) shows a small set of the data. The complete data are available at the text website www.stat.ufl.edu/~aa/cda/cda.html. For now, we use width alone as a predictor. [Table 4.3](#) lists width in centimeters. The sample mean width equals 26.3 and the standard deviation equals 2.1.

Table 4.3 Number of Male Satellites by Female Crab's Characteristics^a

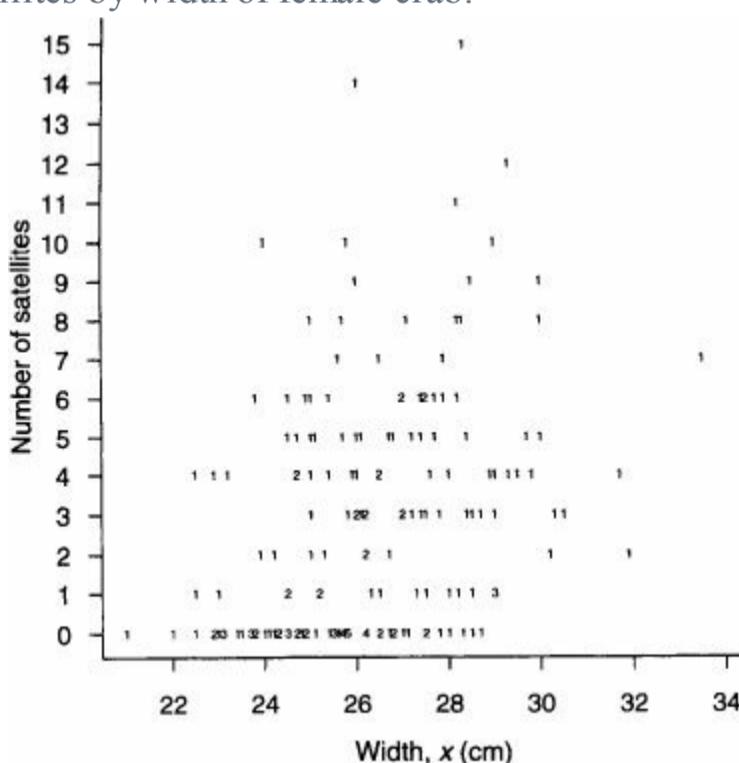
C	S	W	Wt	Sa	C	S	W	Wt	Sa	C	S	W	Wt	Sa
2	3	28.3	3.05	8	3	3	22.5	1.55	0	1	1	26.0	2.30	9
3	3	26.0	2.60	4	2	3	23.8	2.10	0	3	2	24.7	1.90	0
3	3	25.6	2.15	0	3	3	24.3	2.15	0	2	3	25.8	2.65	0
4	2	21.0	1.85	0	2	1	26.0	2.30	14	1	1	27.1	2.95	8

^aC, color (1, light medium; 2, medium; 3, dark medium; 4, dark); S, spine condition (1, both good; 2, one worn or broken; 3, both worn or broken); W, carapace width (cm); Wt, weight (kg); Sa, number of satellites.

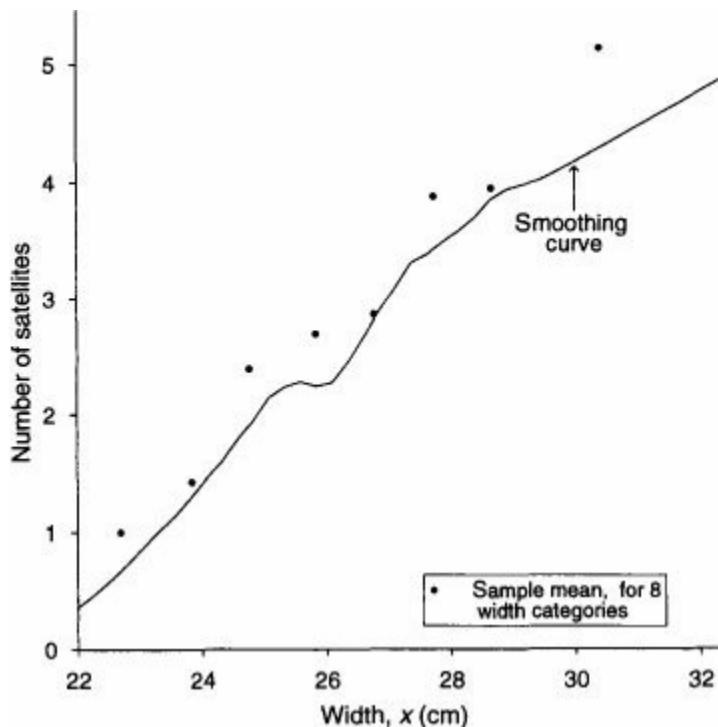
Source: Data courtesy of Jane Brockmann, Zoology Department, University of Florida; study described in *Ethology* 102: 1–21, 1996. The complete data are at the text website.

[Figure 4.3](#) plots the response counts of satellites against width, with numerical symbols indicating the number of observations at each point. The substantial variability makes it difficult to discern a clear trend. To get a clearer picture, we grouped the female crabs into width categories (≤ 23.25 , $23.25\text{--}24.25$, $24.25\text{--}25.25$, $25.25\text{--}26.25$, $26.25\text{--}27.25$, $27.25\text{--}28.25$, $28.25\text{--}29.25$, >29.25) and calculated the sample mean number of satellites for female crabs in each category. [Figure 4.4](#) plots these sample means against the sample mean width for crabs in each category.

[Figure 4.3](#) Number of satellites by width of female crab.



[Figure 4.4](#) Smoothings of horseshoe crab counts.



More sophisticated ways of portraying the trend smooth the data without grouping the width values or assuming a particular functional relationship. [Figure 4.4](#) also shows a smoothed curve based on a semiparametric extension of the GLM (the *generalized additive model*) presented in Section 7.4.9. The sample means and the smoothed curve both show a strong increasing trend. (The means tend to fall above the curve, since the response counts in a category tend to be skewed to the right; the smoothed curve is less susceptible to outlying observations.) The trend seems approximately linear, and we discuss the next models for the ungrouped data for which the mean or the log of the mean is linear in width.

For a female crab, let $\mu(x)$ be the expected number of satellites at width x . From GLM software as shown in the Appendix at the text website, the ML fit of the Poisson loglinear model [\(4.11\)](#) is

$$\log \hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x = -3.305 + 0.164x.$$

The effect $\hat{\beta} = 0.164$ of width is positive, with $SE = 0.020$. The model fitted value at a width level x is an estimated mean number of satellites $\hat{\mu}(x)$. For instance, the fitted value at the mean width of $x = 26.3$ is

$$\hat{\mu}(x) = \exp(\hat{\alpha} + \hat{\beta}x) = \exp[-3.305 + 0.164(26.3)] = 2.74.$$

For this model, $\exp(\hat{\beta}) = \exp(0.164) = 1.18$ is the multiplicative effect on $\mu(x)$ for a 1-cm increase in x . For instance, the fitted value at $x = 27.3 = 26.3 + 1$ is $\exp[-3.305 + 0.164(27.3)] = 3.23$, which equals 1.18×2.74 . A 1-cm increase in width yields an 18% increase in the estimated mean.

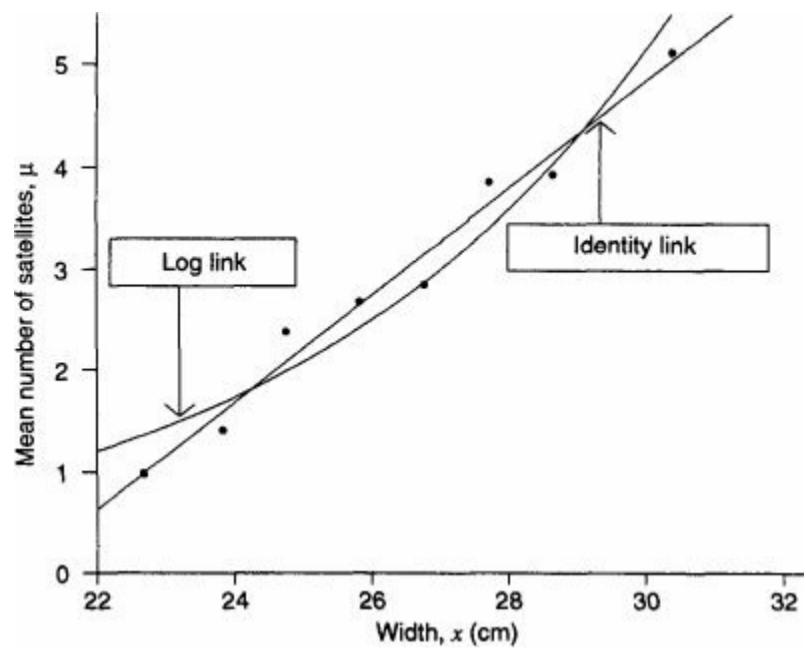
[Figure 4.4](#) shows that $\mu(x)$ may grow approximately linearly with width. This suggests the Poisson GLM with identity link. It has ML fit

$$\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x = -11.53 + 0.55x.$$

This model has an additive rather than a multiplicative effect of x : A 1-cm increase in x has an estimated increase of $\hat{\beta} = 0.55$ in $\hat{\mu}(x)$. The fitted values are positive at all sampled x , and the model describes the effect simply: On the average, about a 2-cm increase in width is associated with an extra satellite.

[Figure 4.5](#) plots $\hat{\mu}(x)$ against width for the models with log link and identity link. Although they diverge somewhat for relatively small and large widths, they provide similar predictions over the width range in which most observations occur. We now study whether either model fits adequately.

[Figure 4.5](#) Estimated mean number of satellites for log and identity links.



4.3.3 Overdispersion for Poisson GLMs

In Section 1.2.4 we noted that count data often show greater variability than the Poisson allows. For the grouped horseshoe crab data, [Table 4.4](#) shows the sample mean and variance for the counts of number of satellites for the female crabs in each width category. The variances are much larger than the means, whereas Poisson distributions have identical mean and variance. The greater variability than predicted by the GLM random component reflects *overdispersion*.

Table 4.4 Sample Mean and Variance of Number of Satellites

Width (cm)	Number of Cases	Number of Satellites	Sample Mean	Sample Variance
<23.25	14	14	1.00	2.77
23.25–24.25	14	20	1.43	8.88
24.25–25.25	28	67	2.39	6.54
25.25–26.25	39	105	2.69	11.38
26.25–27.25	22	63	2.86	6.88
27.25–28.25	24	93	3.87	8.81
28.25–29.25	18	71	3.94	16.88
>29.25	14	72	5.14	8.29

A common cause of overdispersion is subject heterogeneity. For instance, suppose that width, weight, color, and spine condition are the four predictors that affect a female crab's number of satellites. Suppose that Y has a Poisson distribution at each fixed combination of those predictors. Our model uses width alone as a predictor. Crabs having a certain width are then a mixture of crabs of various weights, colors, and spine conditions. Thus, the population of crabs having that width is a mixture of several Poisson populations, each having its own mean for the response. This heterogeneity results in an overall response distribution at that width having greater variation than the Poisson predicts. If the variance equals the mean when all relevant variables are controlled, it exceeds the mean when only one is controlled.

Overdispersion is not an issue in ordinary regression with normally distributed Y , because that distribution has a separate variance parameter to describe variability. For binomial and Poisson distributions, however, the variance is a function of the mean. Overdispersion is common in the modeling of counts. When the model for the mean is correct but the true distribution is not Poisson, the ML estimates of model parameters are still consistent but standard errors are incorrect. We next introduce an extension of the Poisson GLM that has an extra parameter and accounts better for overdispersion. In Section 4.7 we present another approach for this, quasi-likelihood inference.

4.3.4 Negative Binomial GLMs

The *negative binomial distribution* has probability mass function

$$(4.13) \quad f(y; k, \mu) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{k}{\mu+k}\right)^k \left(1 - \frac{k}{\mu+k}\right)^y, \quad y = 0, 1, 2, \dots,$$

where $k > 0$ and $\mu > 0$ are parameters. This distribution results when, given the mean, Y has a Poisson distribution, but the mean itself varies according to a gamma distribution with shape parameter k (Section 14.4).

Notationally, we'll find it simpler to parameterize the negative binomial distribution in terms of μ and $\gamma = 1/k$. Then, Y has

$$E(Y) = \mu, \quad \text{var}(Y) = \mu + \gamma\mu^2.$$

The index $\gamma > 0$ is a type of dispersion parameter. As $\gamma \rightarrow 0$, $\text{var}(Y) \rightarrow \mu$ and the negative binomial distribution converges to the Poisson. Usually, γ is unknown. Estimating it helps summarize the extent of overdispersion. For $k = \gamma$ fixed, we can express (4.13) in natural exponential family form (4.1). Then, a model with negative binomial random component is a GLM. For simplicity, such models let γ be the same constant for all observations but treat it as unknown.

In Section 14.4 we present more detail about negative binomial GLMs. We illustrate the model here for the horseshoe crab data analyzed above with Poisson GLMs. With the identity link and width as predictor, the Poisson GLM has $\hat{\mu} = -11.53 + 0.55x$ ($SE = 0.06$ for β). For the negative binomial GLM, convergence problems are caused by a slightly negative predicted response during the iterative fitting process at the lowest observed width value of 21 cm. Without that observation, the ML fit is $\hat{\mu} = -11.47 + 0.55x$ ($SE = 0.12$). Moreover, $\hat{\gamma} = 1.07$, so at a predicted $\hat{\mu}$, the estimated $\text{var}(Y)$ is roughly $\hat{\mu} + \hat{\mu}^2$, compared to $\hat{\mu}$ for the Poisson GLM. (The fit is similar to that of the *geometric* distribution, which is the special case of the negative binomial with $\gamma = 1.0$.) Although fitted values are similar to the Poisson GLM, the greater SE for β and the greater estimated $\text{var}(Y)$ in the negative binomial model reflect the overdispersion uncaptured with the Poisson GLM. Further improved fit is obtained by allowing “zero-inflation,” permitting a higher fitted count at 0 than the negative binomial model allows.

4.3.5 Poisson Regression for Rates Using Offsets

Often a response count Y_i , has an index t_i , such that its expected value is proportional to t_i . For instance, this index could be an amount of time or a spatial area over which the count is made. Then, the sample rate is y_i/t_i , with expected value μ_i/t_i . With explanatory variables \mathbf{x} , a loglinear model for the expected rate has form

$$(4.14) \log(\mu_i/t_i) = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

This model has equivalent representation

$$\log \mu_i - \log t_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

The adjustment term, $-\log t_i$, to the log link of the mean is called an *offset*. The fit corresponds to using $\log t_i$ as a predictor on the right-hand side and forcing its coefficient to equal 1.0.

For model (4.14), the expected response count satisfies

$$\mu_i = t_i \exp(\alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}).$$

The mean has proportionality constant depending on the value of \mathbf{x}_i . The identity link is also sometimes useful. The model is then

$$\mu_i/t_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \quad \text{or} \quad \mu_i = \alpha t_i + \beta_1 x_{i1} t_i + \cdots + \beta_p x_{ip} t_i.$$

This does not require an offset. It corresponds to an ordinary Poisson GLM using identity link with no intercept and with explanatory variables $t_i, x_{i1}t_i, \dots, x_{ip}t_i$. It provides additive, rather than multiplicative, predictor effects. It is less useful with several predictors, as the fitting process may fail because of a negative fitted count at an \mathbf{x}_i at some step in the iterative process.

4.3.6 Example: Modeling Death Rates for Heart Valve Operations

Laird and Olivier (1981) analyzed patient survival after heart valve replacement operations. A sample of 109 patients were classified by type of heart valve (aortic, mitral) and by age (<55 , ≥ 55). Follow-up observations occurred until the patient died or the study ended. Operations occurred throughout the study period, and follow-up observations covered lengths of time varying from 3 to 97 months. The response was whether the subject died and the follow-up time. For subjects who died, this is the time after the operation until death; for the others, it is the time until the study ended or the subject withdrew from it.

[Table 4.5](#) lists the numbers of deaths during the follow-up period, by valve type and age. These counts are the first layer of a three-way contingency table that classifies valve type, age, and whether died (yes, no). The subjects not tabulated in [Table 4.5](#) were not observed to die. They are *censored*, since we know only a lower bound for how long they lived after the operation. It is inappropriate to analyze that $2 \times 2 \times 2$ table using binary GLMs for the probability of death, since subjects had differing times at risk; it is not sensible to treat a subject who could be observed for 3 months and a subject who could be observed for 97 months as identical trials with the same probability. To use age and valve type as predictors in a model for frequency of death, the proper baseline is not the number of subjects but rather the total time that subjects were at risk. Thus, we model the *rate* of death.

Table 4.5 Data on Heart Valve Replacement Operations

Age		Type of Heart Valve	
		Aortic	Mitral
<55	Number of deaths	4	1
	Time at risk (months)	1259	2082
	Death rate	0.0032	0.0005
55+	Number of deaths	7	9
	Time at risk (months)	1417	1647
	Death rate	0.0049	0.0055

Source: Reprinted with permission, based on data in Laird and Olivier (1981).

The *time at risk* for a subject is their follow-up time of observation. For a given age and valve type, the total time at risk is the sum of the times at risk for all subjects in that cell (those who died and those censored). The sample rate, also shown in that table, divides the number of deaths by total time at risk, in months. For instance, 4 deaths in 1259 months of observation occurred for younger subjects with aortic valve replacement, so their sample rate is $4/1259 = 0.0032$.

We now model effects of age and valve type on the rate. Let Y_{ij} denote the number of deaths for age a_i and valve type v_j , with expected value μ_{ij} . for total time at risk t_{ij} . Given t_{ij} , the expected rate is μ_{ij}/t_{ij} . Let a be an indicator variable for age, with $a_1 = 0$ for the younger age group and $a_2 = 1$ for the older group. Let v be an indicator variable for valve type, with $v_1 = 0$ for aortic and $v_2 = 1$ for mitral. The model

$$(4.15) \log(\mu_{ij}/t_{ij}) = \alpha + \beta_1 a_i + \beta_2 v_j$$

assumes a lack of interaction in the effects of age and valve type.

Using software (as shown at the text website), we treat $\{Y_{ij}\}$ as independent Poisson variates with means $\{\mu_{ij}\}$, conditional on $\{t_{ij}\}$. [Table 4.6](#) presents the fitted death counts and estimated rates. The estimated effects are

Table 4.6 Fit of Poisson Regression Models to [Table 4.5](#) on Heart Valve Operation Deaths

Age		Log Link		Identity Link	
		Aortic	Mitral	Aortic	Mitral
<55	Number of deaths	2.28	2.72	3.16	1.19
	Death rate	0.0018	0.0013	0.0025	0.0006
55+	Number of deaths	8.72	7.28	9.17	7.48
	Death rate	0.0062	0.0044	0.0065	0.0046

$$\hat{\beta}_1 = 1.221 \ (SE = 0.514), \quad \hat{\beta}_2 = -0.330 \ (SE = 0.438).$$

There is evidence of an age effect Given valve type, the estimated death rate for the older age group is $\exp(1.221) = 3.39$ times that for the younger age group. The study contains much censored data. Of the 109 patients, only 21 died during the study period, so both effect estimates are imprecise. Note, though, that the analysis uses all 109 patients through their contributions to the times at risk.

[Table 4.6](#) also shows the fit of the corresponding model with identity link,

$$\mu_{ij} = \alpha t_{ij} + \beta_1 a_i t_{ij} + \beta_2 v_j t_{ij}$$

Substantive conclusions are similar. The estimate $\hat{\beta}_1 = 0.0040 \ (SE = 0.0014)$ represents the estimated difference in death rates between the older and younger age groups for each valve type.

4.3.7 Poisson GLM of Independence in Two-Way Contingency Tables

Poisson loglinear models are also used to model counts in ordinary contingency tables. We illustrate for two-way tables with independent counts $\{Y_{ij}\}$ having Poisson distributions with means $\{\mu_{ij}\}$. Suppose that $\{\mu_{ij}\}$ satisfy

$$\mu_{ij} = \mu \alpha_i \beta_j,$$

where $\{\alpha_i\}$ and $\{\beta_j\}$ are positive constants satisfying $\sum_i \alpha_i = \sum_j \beta_j = 1$. This is a multiplicative model, but a linear predictor for a GLM results using the log link,

$$(4.16) \quad \log \mu_{ij} = \lambda + \alpha_i^* + \beta_j^*,$$

where $\lambda = \log \mu$, $\alpha_i^* = \log \alpha_i$, $\beta_j^* = \log \beta_j$. This Poisson loglinear model has additive main effects of the two classifications but no interaction.

Since the $\{Y_{ij}\}$ are independent, the total sample size $\sum_i \sum_j Y_{ij}$ has a Poisson distribution with mean $\sum_i \sum_j \mu_{ij} = \mu$. Conditional on $\sum_i \sum_j Y_{ij} = n$, the cell counts have a multinomial distribution with probabilities $\{\pi_{ij} = \mu_{ij}/\mu = \alpha_i \beta_j = \pi_{i+} \pi_{+j}\}$. This is *independence* between the two categorical variables. In fact, in Poisson form, independence is the loglinear model (4.16). The inferences conducted in Chapter 3 about independence in two-way contingency tables relate to GLMs, either Poisson loglinear models or corresponding multinomial models that fix n or the row or column totals.

4.4 MOMENTS AND LIKELIHOOD FOR GENERALIZED LINEAR MODELS

Having introduced GLMs for binary and count data, we now turn our attention to the likelihood equations and methods for fitting GLMs. The remainder of this chapter is somewhat technical, providing general results applying to the modeling methods presented in subsequent chapters. Some readers may prefer to skip this material.

4.4.1 The Exponential Dispersion Family

It is helpful to extend the notation for a GLM to handle many distributions that have a second parameter. The random component of the GLM specifies that the N observations (y_1, \dots, y_N) on Y are independent, with probability mass or density function for y_i of form

$$(4.17) \quad f(y_i; \theta_i, \phi) = \exp\{[y_i \theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\}.$$

This is called the *exponential dispersion family* and ϕ is called the *dispersion parameter* (Jørgensen 1987). The parameter θ_i is the *natural parameter*.

When ϕ is known, (4.17) simplifies to the form (4.1) for the natural exponential family, which is

$$f(y_i; \theta_i) = a(\theta_i)b(y_i) \exp[y_i Q(\theta_i)].$$

We identify $Q(\theta)$ here with $\theta/a(\phi)$ in (4.17), $a(\theta)$ with $\exp[-\theta/a(\phi)]$ in (4.17), and $b(y)$ with $\exp[c(y, \phi)]$ in (4.17). The more general formula (4.17) is not needed for one-parameter families such as the binomial and Poisson. Usually, $a(\phi)$ has form $a(\phi) = \phi/\omega_i$ for a known weight ω_i . For instance, when y_i is a mean of n_i independent readings, such as a sample proportion for n_i Bernoulli trials, $\phi = 1$ and $\omega_i = n_i$ (Section 4.4.3).

4.4.2 Mean and Variance Functions for the Random Component

Expressions for $E(Y_i)$ and $\text{var}(Y_i)$ use terms in (4.17). Let $L_i = \log f(y_i; \theta_i, \phi)$ denote the contribution of y_i to the log likelihood, so the log-likelihood function is $L = \sum_i L_i$. From (4.17),

$$(4.18) \quad L_i = [y_i \theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi).$$

Therefore,

$$\partial L_i / \partial \theta_i = [y_i - b'(\theta_i)]/a(\phi), \quad \partial^2 L_i / \partial \theta_i^2 = -b''(\theta_i)/a(\phi),$$

where $b'(\theta_i)$ and $b''(\theta_i)$ denote the first two derivatives of $b(\cdot)$ evaluated at θ_i . We now apply the general likelihood results

$$E\left(\frac{\partial L}{\partial \theta}\right) = 0 \quad \text{and} \quad -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = E\left(\frac{\partial L}{\partial \theta}\right)^2,$$

which hold under regularity conditions satisfied by the exponential family (Cox and Hinkley 1974, Sec. 4.8). From the first formula applied with a single observation, $E[Y_i - b'(\theta_i)]/a(\phi) = 0$, or

$$(4.19) \quad \mu_i = E(Y_i) = b'(\theta_i).$$

From the second formula,

$$b''(\theta_i)/a(\phi) = E[(Y_i - b'(\theta_i))/a(\phi)]^2 = \text{var}(Y_i)/[a(\phi)]^2,$$

so that

$$(4.20) \quad \text{var}(Y_i) = b''(\theta_i)a(\phi).$$

In summary, the function $b(\cdot)$ in (4.17) determines moments of Y_i . This function is called the *cumulant function*, since when $a(\phi) = 1$ its derivatives yield the cumulants of the distribution (Jørgensen 1987).

4.4.3 Mean and Variance Functions for Poisson and Binomial GLMs

We illustrate the mean and variance expressions for Poisson and binomial distributions. When Y_i is Poisson,

$$\begin{aligned} f(y_i; \mu_i) &= \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} = \exp(y_i \log \mu_i - \mu_i - \log y_i!) \\ &= \exp[y_i \theta_i - \exp(\theta_i) - \log y_i!], \end{aligned}$$

where $\theta_i = \log \mu_i$. This has exponential dispersion form (4.17) with $b(\theta_i) = \exp(\theta_i)$, $a(\phi) = 1$, and $c(y_i, \phi) = -\log y_i!$. The natural parameter is $\theta_i = \log \mu_i$. From (4.19) and (4.20),

$$E(Y_i) = b'(\theta_i) = \exp(\theta_i) = \mu_i,$$

$$\text{var}(Y_i) = b''(\theta_i) = \exp(\theta_i) = \mu_i.$$

Next, suppose that $n_i Y_i$ has a $\text{bin}(n_i, \pi_i)$ distribution; that is, here y_i is the sample *proportion* (rather than *number*) of successes, so $E(Y_i) = \pi_i$ does not depend on n_i . Let $\theta_i = \log[\pi_i/(1 - \pi_i)]$. Then, $\pi_i = \exp(\theta_i)/[1 + \exp(\theta_i)]$ and $\log(1 - \pi_i) = -\log[1 + \exp(\theta_i)]$. Extending (4.3), we have the result

$$\begin{aligned} f(y_i; \pi_i, n_i) &= \binom{n_i}{n_i y_i} \pi_i^{n_i y_i} (1 - \pi_i)^{n_i - n_i y_i} \\ (4.21) \quad &= \exp\left[\frac{y_i \theta_i - \log[1 + \exp(\theta_i)]}{1/n_i} + \log\left(\binom{n_i}{n_i y_i}\right)\right]. \end{aligned}$$

This has exponential dispersion form (4.17) with $b(\theta_i) = \log[1 + \exp(\theta_i)]$, $a(\phi) = 1/n_i$, and $c(y_i, \phi) = \log\left(\binom{n_i}{n_i y_i}\right)$. The natural parameter is the logit, $\theta_i = \log[\pi_i/(1 - \pi_i)]$. From (4.19) and (4.20),

$$E(Y_i) = b'(\theta_i) = \exp(\theta_i)/[1 + \exp(\theta_i)] = \pi_i,$$

$$\text{var}(Y_i) = b''(\theta_i) a(\phi) = \exp(\theta_i)/\{[1 + \exp(\theta_i)]^2 n_i\} = \pi_i(1 - \pi_i)/n_i.$$

4.4.4 Systematic Component and Link Function of a GLM

For observation i , the systematic component of a GLM relates parameters $\{\eta_i\}$ to the explanatory variables using a linear predictor

$$\eta_i = \sum_j \beta_j x_{ij}, \quad i = 1, \dots, N.$$

In matrix form,

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

where $\boldsymbol{\eta} = (\eta_1, \dots, \eta_N)^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots)^T$ is the column vector of model parameters. With p parameters in $\boldsymbol{\beta}$, \mathbf{X} is the $N \times p$ matrix of explanatory variable values for the N subjects. In ordinary linear models, \mathbf{X} is called the *design matrix*. It need not refer to an experimental design, however, and the GLM literature calls it the *model matrix*.

The GLM links η_i to $\mu_i = E(Y_i)$ by a link function $g(\cdot)$. Thus, μ_i relates to the explanatory variables by

$$\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij}, \quad i = 1, \dots, N.$$

The link function g for which $g(\mu_i) = \theta_i$ in (4.17) is the *canonical link*. For it, the direct relationship

$$\theta_i = \sum_j \beta_j x_{ij}$$

occurs between the natural parameter and the linear predictor.

4.4.5 Likelihood Equations for a GLM

For N independent observations, from (4.18) the log likelihood is

$$(4.22) \quad L(\beta) = \sum_i L_i = \sum_i \log f(y_i; \theta_i, \phi) = \sum_i \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + \sum_i c(y_i, \phi).$$

The notation $L(\beta)$ reflects the dependence of θ on the model parameters β .

The likelihood equations are

$$\partial L(\beta)/\partial \beta_j = \sum_i \partial L_i/\partial \beta_j = 0$$

for all j . To differentiate the log likelihood (4.22), we use the chain rule,

$$(4.23) \quad \frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}.$$

Since $\partial L_i/\partial \theta_i = [y_i - b'(\theta_i)]/a(\phi)$, and since $\mu_i = b'(\theta_i)$ and $\text{var}(Y_i) = b''(\theta_i)a(\phi)$ from (4.19) and (4.20),

$$\partial L_i/\partial \theta_i = (y_i - \mu_i)/a(\phi), \quad \partial \mu_i/\partial \theta_i = b''(\theta_i) = \text{var}(Y_i)/a(\phi).$$

Also, since $\eta_i = \sum_j \beta_j x_{ij}$

$$\partial \eta_i/\partial \beta_j = x_{ij}.$$

Finally, since $\eta_i = g(\mu_i)$, $\partial \mu_i/\partial \eta_i$ depends on the link function for the model. In summary, substituting into (4.23) gives us

$$(4.24) \quad \frac{\partial L_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a(\phi)} \frac{a(\phi)}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}.$$

Summing over the N observations yields the likelihood equations,

$$(4.25) \quad \sum_{i=1}^N \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 0, 1, 2, \dots$$

Although β does not appear in these equations, it is there implicitly through μ_i , since $\mu_i = g^{-1}(\sum_j \beta_j x_{ij})$. Different link functions yield different sets of equations.

4.4.6 The Key Role of the Mean–Variance Relationship

Interestingly, the likelihood [equations \(4.25\)](#) depend on the distribution of Y_i only through μ_i and $\text{var}(Y_i)$. The variance itself depends on the mean through a particular functional form

$$\text{var}(Y_i) = v(\mu_i)$$

for some function v . For example, $v(\mu_i) = \mu_i$ for the Poisson, $v(\mu_i) = \mu_i(1 - \mu_i)/n_i$ for the binomial proportion, and $v(\mu_i) = \sigma^2$ (i.e., constant) for the normal.

When Y_i has distribution in the natural exponential family, the relationship between the mean and the variance characterizes the distribution (Jørgensen 1987). For instance, if Y_i has distribution in the natural exponential family and if $v(\mu_i) = \mu_i$, then necessarily Y_i has the Poisson distribution.

4.4.7 Likelihood Equations for Binomial GLMs

Suppose that $n_i Y_i$ has a $\text{bin}(n_i, \pi_i)$ distribution. We use the binomial parameterization of Section 4.4.3, so y_i is a sample proportion of successes for n_i trials. The binomial GLM (4.8) for a single predictor extends with several predictors to

$$(4.26) \quad \pi_i = \Phi\left(\sum_j \beta_j x_{ij}\right),$$

where Φ is the standard cdf of some class of continuous distributions. Since $\pi_i = \mu_i = \Phi(\eta_i)$ with $\eta_i = \sum_j \beta_j x_{ij}$

$$\frac{\partial \mu_i}{\partial \eta_i} = \phi(\eta_i) = \phi\left(\sum_j \beta_j x_{ij}\right),$$

where $\phi(u) = d\Phi(u)/du$ [i.e., the pdf corresponding to the cdf Φ , not the dispersion parameter in (4.17)]. Since $\text{var}(Y_i) = \pi_i(1 - \pi_i)/n_i$, the likelihood [equations \(4.25\)](#) simplify to

$$(4.27) \quad \sum_i \frac{n_i(y_i - \pi_i)x_{ij}}{\pi_i(1 - \pi_i)} \phi\left(\sum_j \beta_j x_{ij}\right) = 0,$$

where $\pi_i = \Phi\left(\sum_j \beta_j x_{ij}\right)$.

For the logit link, $\eta_i = \log[\pi_i/(1 - \pi_i)]$, so $\partial \eta_i / \partial \pi_i = 1/[\pi_i(1 - \pi_i)]$ and $\partial \mu_i / \partial \eta_i = \partial \pi_i / \partial \eta_i = \pi_i(1 - \pi_i)$. Then the likelihood [equations \(4.27\)](#) simplify to

$$(4.28) \quad \sum_i n_i(y_i - \pi_i)x_{ij} = 0,$$

where $\pi_i = \Phi\left(\sum_j \beta_j x_{ij}\right)$ with Φ as the standard logistic cdf.

4.4.8 Asymptotic Covariance Matrix of Model Parameter Estimators

The likelihood function for the GLM also determines the asymptotic covariance matrix of the ML estimator $\hat{\beta}$. This matrix is the inverse of the information matrix \mathcal{J} , which has elements $E[-\partial^2 L(\beta)/\partial \beta_h \partial \beta_j]$. To find this, for the contribution L_i to the log likelihood we use the helpful result

$$E\left(\frac{\partial^2 L_i}{\partial \beta_h \partial \beta_j}\right) = -E\left(\frac{\partial L_i}{\partial \beta_h}\right)\left(\frac{\partial L_i}{\partial \beta_j}\right),$$

which holds for exponential families (Cox and Hinkley 1974, Sec. 4.8). Thus,

$$\begin{aligned} E\left(\frac{\partial^2 L_i}{\partial \beta_h \partial \beta_j}\right) &= -E\left[\frac{(Y_i - \mu_i)x_{ih}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} \frac{(Y_i - \mu_i)x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}\right] \quad \text{from (4.24)} \\ &= \frac{-x_{ih}x_{ij}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \end{aligned}$$

Since $L(\beta) = \sum_i L_i$

$$E\left(-\frac{\partial^2 L(\beta)}{\partial \beta_h \partial \beta_j}\right) = \sum_{i=1}^N \frac{x_{ih}x_{ij}}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

Let \mathbf{W} be the diagonal matrix with main-diagonal elements

$$(4.29) \quad w_i = (\partial \mu_i / \partial \eta_i)^2 / \text{var}(Y_i).$$

Then, generalizing from the typical element of the information matrix to the entire matrix,

$$(4.30) \quad \mathcal{J} = \mathbf{X}^T \mathbf{W} \mathbf{X}.$$

Note that the form of \mathbf{W} and hence \mathcal{J} depends on the link function.

The asymptotic covariance matrix of $\hat{\beta}$ is estimated by

$$(4.31) \quad \widehat{\text{cov}}(\hat{\beta}) = \widehat{\mathcal{J}}^{-1} = (\mathbf{X}^T \widehat{\mathbf{W}} \mathbf{X})^{-1},$$

where $\widehat{\mathbf{W}}$ is \mathbf{W} evaluated at $\hat{\beta}$. We'll see an example for Poisson GLMs next and for binomial GLMs in Section 5.5.

4.4.9 Likelihood Equations and $\widehat{\text{cov}}(\hat{\beta})$ for Poisson Loglinear Model

The general Poisson loglinear model (4.4) has the matrix form

$$\log \mu = X\beta.$$

For the log link, $\eta_i = \log \mu_i$, so $\mu_i = \exp(\eta_i)$ and $\partial \mu_i / \partial \eta_i = \exp(\eta_i) = \mu_i$. Since $\text{var}(Y_i) = \mu_i$ the likelihood equations (4.25) simplify to

$$(4.32) \quad \sum_i (y_i - \mu_i)x_{ij} = 0.$$

These equate the sufficient statistics $\sum_i y_i x_{ij}$ for β to their expected values. Also, since

$$w_i = (\partial \mu_i / \partial \eta_i)^2 / \text{var}(Y_i) = \mu_i$$

the estimated covariance matrix (4.31) of $\hat{\beta}$ is $(X^T \hat{W} X)^{-1}$, where \hat{W} is the diagonal matrix with elements of $\hat{\mu}$ on the main diagonal.

4.5 INFERENCE AND MODEL CHECKING FOR GENERALIZED LINEAR MODELS

For most GLMs the likelihood [equations \(4.25\)](#) are nonlinear functions of β . For now, we defer details about solving them for the ML estimator $\hat{\beta}$ and focus instead on using the fit for statistical inference.

The Wald, score, and likelihood-ratio methods introduced in Section 1.3.3 for significance testing and interval estimation apply to any GLM. Likelihood-ratio inference utilizes the *deviance* of the GLM.

4.5.1 Deviance and Goodness of Fit

From Section 4.1.5, the *saturated* GLM has a separate parameter for each observation. It gives a perfect fit. This sounds good, but it is not a helpful model. It does not smooth the data or have the advantages that a simpler model has, such as parsimony. Nonetheless, it serves as a baseline for other models, such as for checking model fit.

A saturated model explains all variation by the systematic component of the model. Let $\tilde{\boldsymbol{\theta}}$ denote the estimate of $\boldsymbol{\theta}$ for the saturated model, corresponding to estimated means $\tilde{\mu}_i = y_i$ for all i . For a particular unsaturated model, denote the corresponding ML estimates by $\hat{\boldsymbol{\theta}}$ and $\hat{\mu}_i$. For maximized log likelihood $L(\hat{\boldsymbol{\mu}}; \mathbf{y})$ for that model and maximized log likelihood $L(\mathbf{y}; \mathbf{y})$ in the saturated case,

$$-2 \log \frac{\text{maximum likelihood for model}}{\text{maximum likelihood for saturated model}} = -2[L(\hat{\boldsymbol{\mu}}; \mathbf{y}) - L(\mathbf{y}; \mathbf{y})]$$

describes lack of fit. It is the likelihood-ratio statistic for testing the null hypothesis that the model holds against the alternative that a more general model holds. From (4.22),

$$\begin{aligned} & -2[L(\hat{\boldsymbol{\mu}}; \mathbf{y}) - L(\mathbf{y}; \mathbf{y})] \\ &= 2 \sum_i [y_i \tilde{\theta}_i - b(\tilde{\theta}_i)]/a(\phi) - 2 \sum_i [y_i \hat{\theta}_i - b(\hat{\theta}_i)]/a(\phi). \end{aligned}$$

When $a(\phi)$ in (4.17) has the form $a(\phi) = \phi/\omega_i$, this statistic equals

$$(4.33) \quad 2 \sum_i \omega_i [y_i (\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]/\phi = D(\mathbf{y}; \hat{\boldsymbol{\mu}})/\phi.$$

This is called the *scaled deviance* and $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ is the *deviance*. The greater the scaled deviance, the poorer the fit. For some GLMs the scaled deviance has an approximate chi-squared distribution.

4.5.2 Deviance for Poisson GLMs

For Poisson GLMs, by Section 4.4.3, $\hat{\theta}_i = \log \hat{\mu}_i$, and $b(\hat{\theta}_i) = \exp(\hat{\theta}_i) = \hat{\mu}_i$. Similarly, $\tilde{\theta}_i = \log y_i$, and $b(\tilde{\theta}_i) = y_i$ for the saturated model. Also $a(\phi) = 1$, so the deviance and scaled deviance (4.33) equal

$$(4.34) \quad D(y; \hat{\mu}) = 2 \sum_i [y_i \log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i].$$

When a model with log link contains an intercept term, the likelihood equation (4.32) implied by that parameter is $\sum_i y_i = \sum_i \hat{\mu}_i$. Then the deviance simplifies to

$$(4.35) \quad D(y; \hat{\mu}) = 2 \sum_i y_i \log(y_i/\hat{\mu}_i).$$

For two-way contingency tables, substituting cell count n_{ij} for y_i and the independence fitted value $\hat{\mu}_{ij}$ for $\hat{\mu}_i$, this reduces to the G^2 statistic (3.11) in Section 3.2.1. For a Poisson or multinomial model applied to a contingency table with a fixed number of cells N , Section 16.3 shows that the deviance has an approximate chi-squared distribution for large $\{\mu_i\}$.

4.5.3 Deviance for Binomial GLMs: Grouped Versus Ungrouped Data

Now consider binomial GLMs with sample proportions $\{y_i\}$ based on $\{n_i\}$ trials. By Section 4.4.3, $\hat{\theta}_i = \log[\hat{\pi}_i/(1 - \hat{\pi}_i)]$ and $b(\hat{\theta}_i) = \log[1 + \exp(\hat{\theta}_i)] = -\log(1 - \hat{\pi}_i)$. Similarly, $\tilde{\theta}_i = \log[y_i/(1 - y_i)]$ and $b(\tilde{\theta}_i) = -\log(1 - y_i)$ for the saturated model. Also, $a(\phi) = 1/n_i$, so $\phi = 1$ and $\omega_i = n_i$. The deviance (4.33) equals

$$\begin{aligned} & 2 \sum_i n_i \left\{ y_i \left(\log \frac{y_i}{1 - y_i} - \log \frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right) + \log(1 - y_i) - \log(1 - \hat{\pi}_i) \right\} \\ &= 2 \sum_i n_i y_i \log \frac{n_i y_i}{n_i - n_i y_i} - 2 \sum_i n_i y_i \log \frac{n_i \hat{\pi}_i}{n_i - n_i \hat{\pi}_i} + 2 \sum_i n_i \log \frac{1 - y_i}{1 - \hat{\pi}_i} \\ &= 2 \sum_i n_i y_i \log \frac{n_i y_i}{n_i \hat{\pi}_i} + 2 \sum_i (n_i - n_i y_i) \log \frac{n_i - n_i y_i}{n_i - n_i \hat{\pi}_i}. \end{aligned}$$

At setting i , $n_i y_i$ is the number of successes and $(n_i - n_i y_i)$ is the number of failures, $i = 1, \dots, N$. Thus, the deviance is a sum over the $2N$ cells of successes and failures and has the same form,

$$(4.36) \quad D(\mathbf{y}; \hat{\mu}) = 2 \sum \text{observed} \times \log(\text{observed}/\text{fitted}),$$

as the deviance (4.35) for Poisson loglinear models with intercept term. With binomial responses, it is possible to construct the data file as expressed here with the counts of successes and failures at each setting for the predictors, or with the individual Bernoulli 0 or 1 observations at the subject level. The deviance differs in the two cases. In the first case the saturated model has a parameter at each setting for the predictors, whereas in the second case it has a parameter for each subject. We refer to these as *grouped data* and *ungrouped data* cases. The approximate chi-squared distribution for the deviance occurs for grouped data but not for ungrouped data (see Exercises 4.5, 4.18, and 5.35). With grouped data, the sample size increases for a fixed number of settings of the predictors and hence a fixed number of parameters for the saturated model.

4.5.4 Likelihood-Ratio Model Comparison Using the Deviances

For a Poisson or binomial model denoted by M , $\phi = 1$, so the deviance (4.33) equals

$$(4.37) \quad D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = -2[L(\hat{\boldsymbol{\mu}}; \mathbf{y}) - L(\mathbf{y}; \mathbf{y})].$$

Consider two models, M_0 with fitted values $\hat{\boldsymbol{\mu}}_0$ and M_1 with fitted values $\hat{\boldsymbol{\mu}}_1$, with M_0 a special case of M_1 . Model M_0 is said to be *nested* within M_1 .

Since M_0 is simpler than M_1 , a smaller set of parameter values satisfies M_0 than satisfies M_1 . Maximizing the log likelihood over a smaller space cannot yield a larger maximum. Thus, $L(\hat{\boldsymbol{\mu}}_0; \mathbf{y}) \leq L(\hat{\boldsymbol{\mu}}_1; \mathbf{y})$, and it follows from (4.37) with the same $L(\mathbf{y}; \mathbf{y})$ for each model that

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}_1) \leq D(\mathbf{y}; \hat{\boldsymbol{\mu}}_0).$$

Simpler models have larger deviances. Assuming that model M_1 holds, the likelihood-ratio test of the hypothesis that M_0 holds uses the test statistic

$$\begin{aligned} & -2[L(\hat{\boldsymbol{\mu}}_0; \mathbf{y}) - L(\hat{\boldsymbol{\mu}}_1; \mathbf{y})] \\ &= -2[L(\hat{\boldsymbol{\mu}}_0; \mathbf{y}) - L(\mathbf{y}; \mathbf{y})] - \{-2[L(\hat{\boldsymbol{\mu}}_1; \mathbf{y}) - L(\mathbf{y}; \mathbf{y})]\} \\ &= D(\mathbf{y}; \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\mu}}_1). \end{aligned}$$

The likelihood-ratio statistic comparing the two models is simply the difference between the deviances. This statistic is large when M_0 fits poorly compared to M_1 .

In fact, since the part in (4.33) involving the saturated model cancels, the difference between deviances,

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\mu}}_1) = 2 \sum_i \omega_i [y_i(\hat{\theta}_{1i} - \hat{\theta}_{0i}) - b(\hat{\theta}_{1i}) + b(\hat{\theta}_{0i})],$$

also has the form of the deviance. Under regularity conditions, this difference has approximately a chi-squared null distribution with df equal to the difference between the numbers of parameters in the two models.

For binomial GLMs and Poisson loglinear GLMs with intercept, from expressions (4.35) and (4.36) for the deviance, the difference in deviances uses the observed counts and the two sets of fitted values in the form

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\mu}}_1) = 2 \sum_i y_i \log(\hat{\mu}_{1i}/\hat{\mu}_{0i}).$$

In fact, Simon (1973) showed that when observations have distribution in the natural exponential family, this equals

$$(4.38) \quad D(\mathbf{y}; \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\mu}}_1) = 2 \sum_i \hat{\mu}_{1i} \log(\hat{\mu}_{1i}/\hat{\mu}_{0i})$$

for GLMs using the canonical link.³ In the rest of this text, we denote this likelihood-ratio statistic for comparing models by $G^2(M_0|M_1)$.

With binomial responses, the test comparing models is the same whether the data file has grouped or ungrouped form. The saturated model differs in the two cases, but its log likelihood cancels when we form the difference between the deviances.

4.5.5 Score Tests for Goodness of Fit and for Model Comparison

For the common GLMs having variance function $\text{var}(Y_i) = v(\mu_i)$ with $\phi = 1$, the score statistic for testing the model fit has the generalized Pearson form (Lovison 2005, Smyth 2003)

$$(4.39) \quad X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}.$$

For Poisson y_i , for which $v(\hat{\mu}_i) = \hat{\mu}_i$, this has the usual Pearson form of

$$\text{Pearson statistic} = \sum (\text{observed} - \text{fitted})^2 / \text{fitted}.$$

When y_i is a binomial proportion based on n_i trials, for which $v(\hat{\mu}_i) = v(\hat{\pi}_i) = \hat{\pi}_i(1 - \hat{\pi}_i)/n_i$, then the X^2 sum (4.39) over the N binomial success observations is identical to a sum over the $2N$ counts of successes and failures that also has this Pearson form (see also Section 6.2.1).

For two nested models, the Pearson difference $X^2(M_0) - X^2(M_1)$ does not have Pearson form. It is not even necessarily nonnegative. A more appropriate generalized Pearson statistic for comparing models is (Lovison 2005, Rao 1961)

$$(4.40) \quad X^2(M_0|M_1) = \sum_i (\hat{\mu}_{1i} - \hat{\mu}_{0i})^2 / v(\hat{\mu}_{0i}).$$

This has the generalized Pearson form with $\{\hat{\mu}_{1i}\}$ in place of $\{y_i\}$. This is not the score statistic for comparing the models unless M_1 is the saturated model. However, for Poisson models with $v(\hat{\mu}_{0i}) = \hat{\mu}_{0i}$ (and corresponding binomial and multinomial GLMs) it is a quadratic approximation for the difference (4.38) between the deviances and has the same null asymptotic behavior.

Let \mathbf{X} be the model matrix for the full model and let $\mathbf{V}(\hat{\mu}_0)$ be the diagonal matrix of estimated variances of the observations under the simpler model. Then, for the canonical link case, Lovison (2005) showed that the score statistic for comparing models has a somewhat different extended Pearson form comparing the two sets of fitted values,

$$(\hat{\mu}_1 - \hat{\mu}_0)^T \mathbf{X} [\mathbf{X}^T \mathbf{V}(\hat{\mu}_0) \mathbf{X}]^{-1} \mathbf{X}^T (\hat{\mu}_1 - \hat{\mu}_0).$$

Lovison also noted that that statistic bounds below $X^2(M_0|M_1)$. Pregibon (1982) gave the score statistic in the more general case. He showed also that the score statistic is a difference between Pearson goodness-of-fit statistics for the models in which the statistic for the full model is evaluated at fitted values that result from the first step of an iterative fitting process that starts at the ML estimates for the reduced model.

4.5.6 Residuals for GLMs

When a GLM fits poorly according to an overall goodness-of-fit test, examination of residuals highlights where the fit is poor. The *Pearson residual* for observation i is

$$(4.41) \quad e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)}}.$$

For it, $\sum_i e_i^2 = X^2$, the generalized Pearson X^2 statistic. In (4.33) let $D(\mathbf{y}; \hat{\mu}) = \sum_i d_i$, where

$$d_i = 2\omega_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)].$$

The *deviance residual* is

$$(4.42) \quad \sqrt{d_i} \times \text{sign}(y_i - \hat{\mu}_i),$$

for which the sum of squares is the deviance.

For instance, for a Poisson GLM, the Pearson residual is

$$e_i = (y_i - \hat{\mu}_i)/\sqrt{\hat{\mu}_i}.$$

Consider the model of independence for two-way contingency tables. For cell count $y_{ij} = n_{ij}$ and independence fitted value μ_{ij} , the Pearson residual has the form (3.13). Then, $\sum_i \sum_j e_{ij}^2$ is the Pearson X^2 chi-squared statistic (3.10), and $\sum_i \sum_j d_{ij} = G^2$, the likelihood-ratio statistic (3.11) for testing independence.

When the model holds, Pearson and deviance residuals are less variable than standard normal because they compare y_i to the fitted mean $\hat{\mu}_i$ rather than the true mean μ_i (e.g., the denominator of (4.41) estimates $[v(\mu_i)]^{1/2} = [\text{var}(Y_i - \mu_i)]^{1/2}$ rather than $[\text{var}(Y_i - \hat{\mu}_i)]^{1/2}$). When $X^2 = \sum_i e_i^2$ has an approximate chi-squared distribution with $\text{df} = v$, X^2 is asymptotically comparable to the sum of squares of v (rather than N) independent standard normal random variables. Thus, when the model holds, $E(\sum_i e_i^2)/N \approx v/N < 1$.

We prefer to use standardized residuals, which divide each raw residual $(y_i - \hat{\mu}_i)$ by its standard error. Let $\mathbf{V} = \mathbf{V}(\boldsymbol{\mu})$ denote the diagonal matrix of variances of the observations. For GLMs we'll see below that the asymptotic covariance matrix of the vector of raw residuals is

$$\text{cov}(\mathbf{y} - \hat{\mu}) = \mathbf{V}^{1/2}[\mathbf{I} - \mathbf{H}_{at}]\mathbf{V}^{1/2},$$

where \mathbf{I} is the identity matrix and \mathbf{H}_{at} is the generalized *hat matrix*,

$$(4.43) \quad \mathbf{H}_{at} = \mathbf{W}^{1/2} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{1/2}.$$

[Recall that \mathbf{W} is the diagonal matrix with elements $w_i = (\partial \mu_i / \partial \eta_i)^2 / \text{var}(Y_i)$.] Let \hat{h}_i denote the estimated diagonal element of \mathbf{H}_{at} for observation i , called its *leverage*. Then, standardizing by dividing $y_i - \hat{\mu}_i$ by its estimated *SE* yields the standardized residual

$$(4.44) \quad r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)(1 - \hat{h}_i)}} = \frac{e_i}{\sqrt{1 - \hat{h}_i}}.$$

For Poisson GLMs, for instance, $r_i = (y_i - \hat{\mu}_i)/\sqrt{\hat{\mu}_i(1 - \hat{h}_i)}$. Pierce and Schafer (1986) presented standardized deviance residuals.

In linear models the hat matrix is so named because $\mathbf{H}_{at} \times \mathbf{y}$ projects the data to the fitted values, $\hat{\mu}$ = “mu-hat.” For GLMs with link function g , a corresponding relation holds for a linearized approximation for $g(\mathbf{y})$, as discussed in Section 4.6.4 and Exercise 4.29. As in ordinary regression, the greater an observation's leverage, the greater its potential influence on the fit. The leverages fall between 0 and 1 and sum to the number of model parameters. Unlike ordinary regression, the hat values depend on the fit as well as the model matrix, and points that have extreme predictor values need not have high leverage.

4.5.7 Covariance Matrices for Fitted Values and Residuals

We found in (4.31) that the asymptotic covariance matrix of $\hat{\beta}$ is $(X^T W X)^{-1}$. Let D denote the diagonal matrix with elements $\partial \mu_i / \partial \eta_i$. Then,

$$W = DV^{-1}D \quad \text{and} \quad V = DW^{-1}D.$$

Since the vector of linear predictor estimated values relates to $\hat{\beta}$ by $\hat{\eta} = X\hat{\beta}$, its asymptotic covariance matrix is $X(X^T W X)^{-1}X^T$. By the delta method, we can obtain the asymptotic covariance matrix of fitted values from this, as

$$\text{cov}(\hat{\mu}) = DX(X^T W X)^{-1}X^T D.$$

As in ordinary linear models, we can exploit the decomposition

$$(y - \mu) = (y - \hat{\mu}) + (\hat{\mu} - \mu),$$

If $(v - \hat{\mu})$ is asymptotically uncorrected with $(\hat{\mu} - \mu)$, then the asymptotic

$$\text{cov}(y - \hat{\mu}) = V - \text{cov}(\hat{\mu}) = DW^{-1}D - DX(X^T W X)^{-1}X^T D.$$

This equals $V^{1/2}[I - H_{\alpha}]V^{1/2}$ for the hat matrix given in (4.43).

So, why is $(y - \hat{\mu})$ asymptotically uncorrected with $(\hat{\mu} - \mu)$, thus generalizing the exact orthogonal decomposition for linear models? One argument⁴ is as follows: An alternative asymptotically unbiased estimator of μ is $\hat{\mu}^* = [\hat{\mu} + L(y - \hat{\mu})]$, for a $N \times N$ matrix of constants L . But such an estimator cannot be asymptotically more efficient than the ML estimator $\hat{\mu}$. Let $C = \text{cov}(y - \hat{\mu}, \hat{\mu})$ and consider the case $L = -C[\text{cov}(y - \hat{\mu})]^{-1}$. Then, direct calculation shows that the asymptotic covariance matrix of $\hat{\mu}^*$ is

$$\text{cov}(\hat{\mu}^*) = \text{cov}(\hat{\mu}) - C[\text{cov}(y - \hat{\mu})]^{-1}C^T.$$

But this gives the contradiction that $\hat{\mu}^*$ is asymptotically more efficient than $\hat{\mu}$, unless $C = 0$.

4.5.8 The Bayesian Approach for GLMs

There is by now an enormous literature on the Bayesian approach to inference using GLMs, and many books that survey the Bayesian approach spend considerable time on GLMs. For instance, see Dey et al. (2000) and Christensen et al. (2010).

In this book, we'll show some details about the Bayesian approach as we present the various important GLMs for categorical data. In particular, Section 7.2 presents Bayesian methods for binomial regression models. Section 8.6 presents them for multinomial models, and Section 10.7 presents them for Poisson loglinear models. A couple of general results for GLMs are that (1) model parameters in models for categorical data are more commonly treated with normal prior distributions than conjugate priors, and (2) the Jeffreys prior is improper for most GLMs except for binary regression models (Ibrahim and Laud 1991).

4.6 FITTING GENERALIZED LINEAR MODELS

How do we find the ML estimators $\hat{\beta}$ of GLM parameters? The likelihood [equations \(4.25\)](#) are usually nonlinear in β . We describe a general-purpose iterative method for solving nonlinear equations and apply it in two ways to determine the maximum of a likelihood function.

4.6.1 Newton–Raphson Method

The *Newton–Raphson method* is an iterative method for solving nonlinear equations, such as equations whose solution determines the point at which a function takes its maximum. It begins with an initial guess for the solution. It obtains a second guess by approximating the function to be maximized in a neighborhood of the initial guess by a second-degree polynomial and then finding the location of that polynomial's maximum value. It then approximates the function in a neighborhood of the second guess by another second-degree polynomial, and the third guess is the location of its maximum. In this manner, the method generates a sequence of guesses. These converge to the location of the maximum when the function is suitable and/or the initial guess is good.

In more detail, here's how Newton–Raphson determines the value $\hat{\beta}$ at which a function $L(\beta)$ is maximized. Let $\mathbf{u}^T = (\partial L(\beta)/\partial \beta_0, \partial L(\beta)/\partial \beta_1, \dots)$. Let \mathbf{H} denote the matrix having entries $h_{ab} = \partial^2 L(\beta)/\partial \beta_a \partial \beta_b$ called the *Hessian matrix*. Let $\mathbf{u}^{(t)}$ and $\mathbf{H}^{(t)}$ be \mathbf{u} and \mathbf{H} evaluated at $\beta^{(t)}$, the guess t for $\hat{\beta}$. Step t in the iterative process ($t = 0, 1, 2, \dots$) approximates $L(\beta)$ near $\beta^{(t)}$ by the terms up to second order in its Taylor series expansion,

$$L(\beta) \approx L(\beta^{(t)}) + \mathbf{u}^{(t)T}(\beta - \beta^{(t)}) + \left(\frac{1}{2}\right)(\beta - \beta^{(t)})^T \mathbf{H}^{(t)}(\beta - \beta^{(t)}).$$

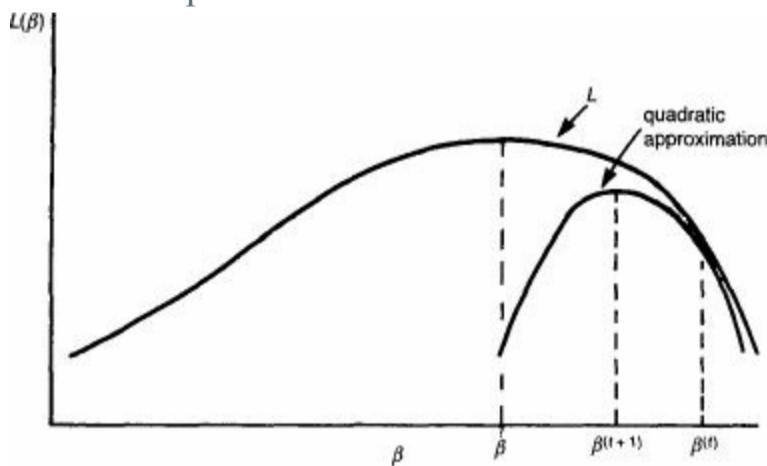
Solving $\partial L(\beta)/\partial \beta \approx \mathbf{u}^{(t)} + \mathbf{H}^{(t)}(\beta - \beta^{(t)}) = \mathbf{0}$ for β yields the next guess. That guess can be expressed as

$$(4.45) \quad \beta^{(t+1)} = \beta^{(t)} - (\mathbf{H}^{(t)})^{-1} \mathbf{u}^{(t)},$$

assuming that $\mathbf{H}^{(t)}$ is nonsingular. (Computing routines use standard methods for solving the linear equations rather than explicitly calculating the inverse.)

Iterations proceed until changes in $L(\beta^{(t)})$ between successive cycles are sufficiently small. The ML estimator is the limit of $\beta^{(t)}$ as $t \rightarrow \infty$; however, this need not happen if $L(\beta)$ has other local maxima at which the derivative of $L(\beta)$ equals 0. In that case, a good initial estimate is crucial. To help understand the Newton–Raphson method, work through these steps when β has a single element (Exercise 4.30). Then, [Figure 4.6](#) illustrates a cycle of the method, showing the parabolic (second-order) approximation at a given step.

[Figure 4.6](#) A cycle of the Newton–Raphson method.



The convergence of $\beta^{(t)}$ to $\hat{\beta}$ for the Newton–Raphson method is usually fast. For large t , the convergence satisfies, for each j ,

$$|\beta_j^{(t+1)} - \hat{\beta}_j| \leq c |\beta_j^{(t)} - \hat{\beta}_j|^2 \quad \text{for some } c > 0$$

and is referred to as *second-order*. This implies that the number of correct decimal places in the approximation roughly doubles after sufficiently many iterations. In practice, it often takes relatively few iterations for satisfactory convergence.

For many GLMs, including Poisson models with log link and binary models with logit link, with full-rank model matrix the Hessian is negative definite and the log likelihood is a strictly concave function. Then ML estimates of model parameters exist and are unique under quite general conditions (Wedderburn 1976).

4.6.2 Fisher Scoring Method

Fisher scoring is an alternative iterative method for solving likelihood equations. It resembles the Newton–Raphson method, the distinction being with the Hessian matrix. Fisher scoring uses the *expected value* of this matrix, called the *expected information*, whereas Newton–Raphson uses the Hessian matrix itself, called the *observed information*.

Let $\mathcal{J}^{(t)}$ denote the approximation t for the ML estimate of the expected information matrix; that is, $\mathcal{J}^{(t)}$ has elements $-E(\partial^2 L(\beta)/\partial\beta_a \partial\beta_b)$, evaluated at $\beta^{(t)}$. The formula for Fisher scoring is

$$\beta^{(t+1)} = \beta^{(t)} + (\mathcal{J}^{(t)})^{-1} u^{(t)}$$

or

$$(4.46) \quad \mathcal{J}^{(t)} \beta^{(t+1)} = \mathcal{J}^{(t)} \beta^{(t)} + u^{(t)}.$$

Formula (4.30) showed that $\mathcal{J} = \mathbf{X}^T \mathbf{W} \mathbf{X}$, where \mathbf{W} is the diagonal matrix with main-diagonal elements $w_i = (\partial\mu_i/\partial\eta_i)^2/\text{var}(Y_i)$. Similarly, $\mathcal{J}^{(t)} = \mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X}$, where $\mathbf{W}^{(t)}$ is \mathbf{W} evaluated at $\beta^{(t)}$. The estimated asymptotic covariance matrix \mathcal{J}^{-1} of β [see (4.31)] occurs as a by-product of this algorithm as $(\mathcal{J}^{(t)})^{-1}$ for t at which convergence is adequate. From (4.25), for both Fisher scoring and Newton–Raphson, the *score function* u has elements

$$(4.47) \quad u_j = \frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^N \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}.$$

Using the matrix $\mathbf{D} = \text{diag}\{\partial\mu_i/\partial\eta_i\}$ introduced in Section 4.5.7, we see that the GLM likelihood equations can be expressed as

$$(4.48) \quad u = \mathbf{X}^T \mathbf{W} \mathbf{D}^{-1} (\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}.$$

For GLMs with a canonical link, we'll see (Section 4.6.5) that the observed and expected information are the same. For noncanonical link models, Fisher scoring has the advantages that it produces the asymptotic covariance matrix as a by-product, the expected information is necessarily nonnegative definite, and as seen next, it is closely related to weighted least-squares methods for ordinary linear models. However, it need not have second-order convergence, and for complex models the observed information is often easier to calculate. Efron and Hinkley (1978), developing arguments of R. A. Fisher, gave reasons for preferring observed information. They argued that its variance estimates better approximate a relevant conditional variance (conditional on statistics not relevant to the parameter being estimated), it is “closer to the data,” and it tends to agree more closely with Bayesian analyses.

4.6.3 Newton–Raphson and Fisher Scoring for Binary Data

In the next three chapters we use the Newton–Raphson and Fisher scoring methods for binary regression models. For now, we illustrate them with a simpler problem for which we know the answer, maximizing the log likelihood based on an observation y from a $\text{bin}(n, \pi)$ distribution.

From Section 1.3.2, the first two derivatives of $L(\pi) = y \log \pi + (n - y) \log(1 - \pi)$ are

$$u = (y - n\pi)/\pi(1 - \pi), \quad H = -[y/\pi^2 + (n - y)/(1 - \pi)^2].$$

Each Newton–Raphson step has the form

$$\pi^{(t+1)} = \pi^{(t)} + \left[\frac{y}{(\pi^{(t)})^2} + \frac{n - y}{(1 - \pi^{(t)})^2} \right]^{-1} \frac{y - n\pi^{(t)}}{\pi^{(t)}(1 - \pi^{(t)})}.$$

This adjusts $\pi^{(t)}$ up if $y/n > \pi^{(t)}$ and down if $y/n < \pi^{(t)}$. For instance, with $\pi^{(0)} = \frac{1}{2}$, you can check that $\pi^{(1)} = y/n$. When $\pi^{(t)} = y/n$, no adjustment occurs and $\pi^{(t+1)} = y/n$, which is the correct answer for $\hat{\pi}$. For starting values other than 1/2, adequate convergence usually takes just a few more iterations.

From Section 1.3.2, the information is $n/[\pi(1 - \pi)]$. A step of Fisher scoring gives

$$\begin{aligned} \pi^{(t+1)} &= \pi^{(t)} + \left[\frac{n}{\pi^{(t)}(1 - \pi^{(t)})} \right]^{-1} \frac{y - n\pi^{(t)}}{\pi^{(t)}(1 - \pi^{(t)})} \\ &= \pi^{(t)} + \frac{y - n\pi^{(t)}}{n} = \frac{y}{n}. \end{aligned}$$

This gives the correct answer for $\hat{\pi}$ after a single iteration and stays at that value for successive iterations.

4.6.4 ML as Iterative Reweighted Least Squares

A relation exists between *weighted least-squares estimation* and using Fisher scoring to find ML estimates. We refer here to the general linear model of form

$$z = \mathbf{X}\beta + \epsilon.$$

When the covariance matrix of ϵ is V , the weighted least-squares (WLS) estimator of β is

$$(\mathbf{X}^T V^{-1} \mathbf{X})^{-1} \mathbf{X}^T V^{-1} z.$$

In practice, V itself must usually be estimated to use this formula.

From $\mathcal{J} = \mathbf{X}^T \mathbf{W} \mathbf{X}$ and expression (4.48) for u , it follows that, in (4.46),

$$\begin{aligned} \mathcal{J}^{(t)} \beta^{(t)} + u^{(t)} &= (\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X}) \beta^{(t)} + \mathbf{X}^T \mathbf{W}^{(t)} (\mathbf{D}^{(t)})^{-1} (y - \mu^{(t)}) \\ &= \mathbf{X}^T \mathbf{W}^{(t)} [\mathbf{X} \beta^{(t)} + (\mathbf{D}^{(t)})^{-1} (y - \mu^{(t)})] = \mathbf{X}^T \mathbf{W}^{(t)} z^{(t)}, \end{aligned}$$

where $z^{(t)}$ has elements

$$z_i^{(t)} = \sum_j x_{ij} \beta_j^{(t)} + (y_i - \mu_i^{(t)}) \frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}} = \eta_i^{(t)} + (y_i - \mu_i^{(t)}) \frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}}.$$

Equations (4.46) for Fisher scoring then have the form

$$(\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X}) \beta^{(t+1)} = \mathbf{X}^T \mathbf{W}^{(t)} z^{(t)}.$$

These are the normal equations for using weighted least squares to fit a linear model for a response variable $z^{(t)}$, when the model matrix is \mathbf{X} and the inverse of the covariance matrix is $\mathbf{W}^{(t)}$. The equations have solution

$$\beta^{(t+1)} = (\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(t)} z^{(t)}.$$

The vector $z^{(t)}$ in this formulation is an estimated linearized form of the link function g , evaluated at y ,

$$(4.49) \quad g(y_i) \approx g(\mu_i^{(t)}) + (y_i - \mu_i^{(t)}) g'(\mu_i^{(t)}) = \eta_i^{(t)} + (y_i - \mu_i^{(t)}) \frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}} = z_i^{(t)}.$$

This *adjusted* (or “working”) *response variable* z has element i approximated by $z_i^{(t)}$ for cycle t of the iterative scheme. That cycle regresses $z^{(t)}$ on \mathbf{X} with weight (i.e., inverse covariance) $\mathbf{W}^{(t)}$ to obtain a new estimate $\beta^{(t+1)}$. This estimate yields a new linear predictor value $\eta^{(t+1)} = \mathbf{X} \beta^{(t+1)}$ and a new adjusted response value $z^{(t+1)}$ for the next cycle. The ML estimator results from iterative use of weighted least squares, in which the weight matrix changes at each cycle. The process is called *iterative reweighted least squares*. The weight matrix \mathbf{W} used in $\text{cov}(\beta)$ [see (4.31)], in the hat matrix (4.43), and in Fisher scoring is the inverse covariance matrix of the linearized form $z = \mathbf{X}\beta + \mathbf{D}^{-1}(y - \mu)$ of $g(y)$.

At convergence,

$$\hat{\beta} = (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{W}} \hat{z},$$

for the estimated linearized response $\hat{z} = \mathbf{X} \hat{\beta} + \hat{\mathbf{D}}^{-1}(y - \hat{\mu})$. Since

$$\hat{\eta} = \mathbf{X} \hat{\beta} = \mathbf{X} (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{W}} \hat{z},$$

$\mathbf{X} (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{W}} = \hat{\mathbf{W}}^{-1/2} (\hat{\mathbf{H}}_{at}) \hat{\mathbf{W}}^{1/2}$ is a sort of asymmetric projection adaptation of the hat matrix shown in (4.43). Tutz (2011, Sec. 3.10) noted the alternative asymmetric projection,

$$\hat{\mu} - \mu \approx \mathbf{V}^{1/2} (\mathbf{H}_{at}) \mathbf{V}^{-1/2} (y - \mu).$$

A simple way to begin the iterative process uses the data y as the initial estimate of μ . This determines the first estimate of the weight matrix \mathbf{W} and hence the initial estimate of β . It may be necessary to alter some observations slightly for this first cycle only so that $g(y)$, the initial value of z , is finite. For instance, when g is the log link applied to counts, a count of $y_i = 0$ is problematic, so we could set $y_i = \frac{1}{2}$. This is not a problem with the model itself, since the log applies to the mean, and fitted means are usually strictly positive in successive iterations.

4.6.5 Simplifications for Canonical Link Functions

Certain simplifications result with GLMs using the canonical link function. For that link,

$$\eta_i = \theta_i = \sum_j \beta_j x_{ij}.$$

Often, $a(\phi)$ in the density or mass function (4.17) is identical for all observations, such as for Poisson GLMs [$a(\phi) = 1$] and binomial GLMs with each $n_i = 1$ [for which $a(\phi) = 1/n_i = 1$]. Then the part of the log likelihood (4.22) involving both parameters and data is $\sum_i y_i \theta_i$, which simplifies to

$$\sum_i y_i \left(\sum_j \beta_j x_{ij} \right) = \sum_j \beta_j \left(\sum_i y_i x_{ij} \right).$$

Sufficient statistics for estimating β in the GLM are then

$$\sum_i y_i x_{ij}, \quad j = 0, 1, 2, \dots$$

For the canonical link,

$$\partial \mu_i / \partial \eta_i = \partial \mu_i / \partial \theta_i = \partial b'(\theta_i) / \partial \theta_i = b''(\theta_i).$$

Since $\text{var}(Y_i) = b''(\theta_i)a(\phi)$, the contribution (4.24) to the likelihood equation for β_j simplifies to

$$(4.50) \quad \frac{\partial L_i}{\partial \beta_j} = \frac{y_i - \mu_i}{\text{var}(Y_i)} b''(\theta_i) x_{ij} = \frac{(y_i - \mu_i) x_{ij}}{a(\phi)}.$$

When $a(\phi)$ is identical for all observations, the likelihood equations are

$$(4.51) \quad \sum_i x_{ij} y_i = \sum_i x_{ij} \mu_i, \quad j = 0, 1, 2, \dots$$

This equation illustrates a fundamental result: *For GLMs with canonical link, the likelihood equations equate the sufficient statistics for the model parameters to their expected values.* For a normal distribution with identity link, these are the *normal equations*. We obtained these for Poisson loglinear models in (4.32) and for binomial logistic regression models (when each $n_i = 1$) in (4.28).

From expression (4.50) for $\partial L_i / \partial \beta_j$, with the canonical link the second derivatives of the log likelihood have components

$$\frac{\partial^2 L_i}{\partial \beta_j \partial \beta_h} = -\frac{x_{ij}}{a(\phi)} \left(\frac{\partial \mu_i}{\partial \beta_h} \right).$$

This does not depend on the observation y_i , so

$$\partial^2 L(\beta) / \partial \beta_h \partial \beta_j = E[\partial^2 L(\beta) / \partial \beta_h \partial \beta_j].$$

That is, $H = -J$, and the Newton–Raphson and Fisher scoring algorithms are identical for canonical link models (Nelder and Wedderburn 1972).

4.7 QUASI-LIKELIHOOD AND GENERALIZED LINEAR MODELS

As noted in Section 4.4.5, the likelihood equations

$$(4.52) \quad \sum_{i=1}^N \frac{(y_i - \mu_i)x_{ij}}{v(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_j} \right) = 0, \quad j = 0, 1, \dots, p,$$

for a GLM depend on the assumed distribution for Y_i only through μ_i and $v(\mu_i)$. The choice of distribution determines the mean–variance relationship $v(\mu_i)$.

4.7.1 Mean–Variance Relationship Determines Quasi-likelihood Estimates

Wedderburn (1974) proposed an alternative approach, *quasi-likelihood estimation*, which assumes only a mean–variance relationship rather than a specific distribution for Y_i . It has a link function and linear predictor of the usual GLM form, but instead of assuming a distributional type for Y_i it assumes only

$$\text{var}(Y_i) = v(\mu_i)$$

for some chosen variance function v . The equations that determine quasi-likelihood estimates are the same as the likelihood [equations \(4.52\)](#) for GLMs. They are not likelihood equations, however, without the additional assumption that $\{Y_i\}$ has distribution in the natural exponential family.

To illustrate, suppose we assume that the $\{Y_i\}$ are independent with

$$v(\mu_i) = \mu_i.$$

The quasi-likelihood (QL) estimates are the solution of [\(4.52\)](#) with $v(\mu_i)$ replaced by μ_i . Under the additional assumption that $\{Y_i\}$ have distribution in the natural exponential family, these estimates are also ML estimates. That case is simply the Poisson distribution. Thus, for $v(\mu) = \mu$, quasi-likelihood estimates are also ML estimates when the random component has a Poisson distribution.

Wedderburn suggested using the estimating [equations \(4.52\)](#) for *any* variance function, even if it does not occur for a member of the natural exponential family. In fact, the purpose of the quasi-likelihood method was to encompass a greater variety of cases, such as discussed next. The QL estimates have asymptotic covariance matrix of the same form [\(4.31\)](#) as in GLMs, namely, $(\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1}$ with $w_i = (\partial \mu_i / \partial \eta_i)^2 / \text{var}(Y_i)$.

4.7.2 Overdispersion for Poisson GLMs and Quasi-likelihood

For count data, we've seen (Section 4.3.3) that the Poisson assumption is often unrealistic because of overdispersion — the variance exceeds the mean. This suggests an alternative to a Poisson GLM in which the mean-variance relationship has the form

$$v(\mu_i) = \phi\mu_i$$

for some constant ϕ . The case $\phi > 1$ represents overdispersion for the Poisson model.

In the estimating [equations \(4.52\)](#) with $v(\mu_i) = \phi\mu_i, \phi$ drops out. Thus, the equations are identical to likelihood equations for Poisson models, and model parameter estimates are also identical. Also,

$$w_i = (\partial\mu_i/\partial\eta_i)^2/\text{var}(Y_i) = (\partial\mu_i/\partial\eta_i)^2/\phi\mu_i,$$

so the estimated $\text{cov}(\hat{\beta}) = (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1}$ is ϕ times that for the Poisson model.

When a variance function has the form $v(\mu_i) = \phi v^*(\mu_i)$, usually ϕ is also unknown. However, ϕ is not in the estimating equations. Let $X^2 = \sum_i (y_i - \hat{\mu}_i)^2/v^*(\hat{\mu}_i)$ a generalized Pearson statistic for the simpler model with $\phi = 1$. When X^2/ϕ is approximately chi-squared or when μ_i is approximately linear in β with $v^*(\hat{\mu}_i)$ close to $v^*(\mu_i)$, then $E(X^2/\phi) \approx N - p$, the number of observations minus the number of model parameters p . Hence, $E[X^2/(N - p)] \approx \phi$. Using the motivation of moment estimation, Wedderburn (1974) suggested taking $\hat{\phi} = X^2/(N - p)$ as the estimated multiple of the covariance matrix.

In summary, this quasi-likelihood approach for count data is simple: Fit the ordinary Poisson model and use its p parameter estimates. Multiply the ordinary standard error estimates by $\sqrt{X^2/(N - p)}$.

We illustrate for the horseshoe crab data analyzed with Poisson GLMs in Section 4.3.2. With the log link, the fit using the crab's carapace width to predict the number of satellites was $\log \hat{\mu} = -3.305 + 0.164x$, with $SE = 0.020$ for $\beta = 0.164$. To improve the adequacy of using a chi-squared statistic to summarize fit, we use the satellite totals and fit for all female crabs at a given width, to increase the counts and fitted values relative to those for individual female crabs. The $N = 66$ distinct width levels each have a total count y_i for the number of satellites and a fitted total $\hat{\mu}$. The Pearson statistic comparing these is $X^2 = 174.3$. The quasi-likelihood adjustment for standard errors equals $\sqrt{174.3/(66 - 2)} = 1.65$. Thus, $SE = 1.65(0.020) = 0.033$ is a more plausible standard error for $\beta = 0.164$ in this prediction equation.

Alternative ways of handling overdispersion include mixture models that allow heterogeneity in the mean at fixed settings of predictors (Chapter 14). For count data these include Poisson GLMs having random effects and negative binomial GLMs that result when a Poisson parameter itself has a gamma distribution.

4.7.3 Overdispersion for Binomial GLMs and Quasi-likelihood

The quasi-likelihood approach can also handle overdispersion for counts based on grouped binary data. Suppose y_i is the sample proportion from n_i Bernoulli trials with parameter $\pi_i, i = 1, \dots, N$. The $\{y_i\}$ may exhibit more variability than the binomial allows because of heterogeneity, with observations at a particular setting of explanatory variables having success probabilities that vary according to values of unobserved variables. Or, extra variability could occur because the Bernoulli trials at each i are positively correlated (Section 14.3.3).

With binomial sampling (i.e., independent, identical trials), $E(Y_i) = \pi_i$ and $\text{var}(Y_i) = \pi_i(1 - \pi_i)/n_i$. A simple quasi-likelihood approach uses the alternative variance function

$$(4.53) \quad v(\pi_i) = \phi\pi_i(1 - \pi_i)/n_i.$$

Overdispersion occurs when $\phi > 1$. The quasi-likelihood estimates are the same as ML estimates for the binomial model, since ϕ drops out of the estimating [equations \(4.52\)](#). As in the overdispersed Poisson case, ϕ enters the denominator of w_i . Thus, the asymptotic covariance matrix multiplies by ϕ , and standard errors multiply by $\sqrt{\phi}$. An estimate of ϕ using the χ^2 fit statistic for the ordinary binomial model with p parameters is $\chi^2/(N - p)$ (Finney 1947).

Methods like these that use estimates from ordinary models but inflate their standard errors are appropriate only if the model chosen describes well the structural relationship between the mean of Y and the predictors. If a large goodness-of-fit statistic is due to some other type of lack of fit, such as failing to include a relevant interaction term, making an adjustment for overdispersion will not address the inadequacy.

For counts with binary data, alternative mechanisms for handling overdispersion include mixture models such as binomial GLMs with random effects (Section 13.3) and models for which a binomial parameter itself has a beta distribution (Section 14.3). These are preferable, because they correspond to an actual model. By contrast, although the approach using the variance formula $v(\pi_i) = \phi\pi_i(1 - \pi_i)/n_i$, has the advantage of simplicity, it has a structural problem when $n_i = 1$: Necessarily $v(\pi_i) = \pi_i(1 - \pi_i)$ for ungrouped binary data, and only $\phi = 1$ then makes sense.

4.7.4 Example: Teratology Overdispersion

Teratology is the study of abnormalities of physiological development. Some teratology experiments investigate effects of dietary regimens or chemical agents on the fetal development of rats in a laboratory setting. [Table 4.7](#) shows results from one such study (Moore and Tsiatis 1991). Female rats on iron-deficient diets were assigned to four groups. Rats in group 1 were given placebo injections, and rats in other groups were given injections of an iron supplement; this was done weekly in group 4, only on days 7 and 10 in group 2, and only on days 0 and 7 in group 3. The 58 rats were made pregnant, sacrificed after three weeks, and then the total number of dead fetuses was counted in each litter. Due to unmeasured covariates and genetic variability the probability of death may vary from litter to litter within a particular treatment group.

Table 4.7 Response Counts of (Litter Size, Number Dead) for 58 Litters of Rats in Low-Iron Teratology Study

Group 1: Untreated (low iron)
(10,1) (11,4) (12,9) (4,4) (10, 10) (11,9) (9,9) (11,11) (10,10) (10,7) (12,12) (10,9) (8, 8) (11,9) (6,4) (9,7) (14,14) (12,7) (11,9) (13, 8) (14, 5) (10, 10) (12,10) (13,8) (10,10) (14,3) (13,13) (4, 3) (8, 8) (13, 5) (12,12)
Group 2: Injections days 7 and 10
(10,1) (3,1) (13,1) (12,0) (14,4) (9,2) (13,2) (16,1) (11,0) (4,0) (1,0) (12,0)
Group 3: Injections days 0 and 7
(8,0) (11,1) (14,0) (14,1) (11,0)
Group 4: Injections weekly
(3,0) (13,0) (9,2) (17,2) (15,0) (2,0) (14, 1) (8,0) (6,0) (17,0)

Source: Moore and Tsiatis (1991).

Let $y_{i(g)}$ denote the proportion dead of the $n_{i(g)}$ fetuses in litter i in treatment group g . Let $\pi_{i(g)}$ denote the probability of death for a fetus in that litter. Consider the model that treats $n_{i(g)}y_{i(g)}$ as a $\text{bin}(n_{i(g)}, \pi_{i(g)})$ variate, where

$$\pi_{i(g)} = \pi_g, \quad g = 1, 2, 3, 4.$$

That is, the model treats all litters in a particular group g as having the same probability of death π_g . The ML fit has estimate $\hat{\pi}_g$ equal to the overall sample proportion of deaths for all fetuses from litters in that group. These equal $\hat{\pi}_1 = 0.758$ ($SE = 0.024$), $\hat{\pi}_2 = 0.102$ ($SE = 0.028$), $\hat{\pi}_3 = 0.034$ ($SE = 0.024$), and $\hat{\pi}_4 = 0.048$ ($SE = 0.021$), where for group g , $SE = \sqrt{\hat{\pi}_g(1 - \hat{\pi}_g)/(\sum_i n_{i(g)})}$. The estimated probability of death is considerably higher for the placebo group.

For litter i in group g , $n_{i(g)}\hat{\pi}_g$ is a fitted number of deaths and $n_{i(g)}(1 - \hat{\pi}_g)$ is a fitted number of nondeaths. Comparing these fitted values with the observed counts of deaths and nondeaths in the $N = 58$ litters using the Pearson statistic gives $X^2 = 154.7$ with $df = 58 - 4 = 54$. There is considerable evidence of overdispersion. With the quasi-likelihood approach, $\{\hat{\pi}_g\}$ are the same as the binomial ML estimates; however, $\hat{\phi} = X^2/(N - p) = 154.7/(58 - 4) = 2.86$, so standard errors multiply by $\hat{\phi}^{1/2} = 1.69$.

Even with this adjustment for overdispersion, strong evidence remains that the probability of death is substantially higher for the placebo group. For instance, a 95% confidence interval for $\pi_1 - \pi_2$ is

$$(0.758 - 0.102) \pm 1.96(1.69)\sqrt{(0.024)^2 + (0.028)^2} \quad \text{or} \quad (0.54, 0.78).$$

This is quite a bit wider than the Wald interval of (0.59, 0.73) for comparing independent proportions, which ignores the overdispersion.

NOTES

Section 4.1: The Generalized Linear Model

4.1 Exponential families: Distribution (4.1) is called a *natural* (or *linear*) exponential family to distinguish it from a more general exponential family that replaces y by $r(y)$ in the exponential term. For other generalizations, see Jørgensen (1983, 1987). Books on GLMs and related models include Aitkin et al. (2009), Fahrmeir and Tutz (2001), Lee et al. (2006), McCullagh and Nelder (1989), and McCulloch et al. (2008).

Section 4.3: Generalized Linear Models for Counts and Rates

4.2 Poisson GLMs: For further discussion of Poisson regression and related models for count data, see Breslow (1984, 1990), Cameron and Trivedi (1998), Frome (1983), Hinde (1982), Lawless (1987), and Seeber (2005) and references therein.

4.3 Rates/survival: Consider a contingency table for rate data (such as Table 4.10 below) in which one dimension is a discrete time scale. Holford (1980) and Laird and Olivier (1981) showed that Poisson loglinear models and likelihoods for this table are equivalent to loglinear hazard models and likelihoods that assume piecewise exponential hazards for the survival times. For short time intervals, this approach is essentially nonparametric and is a discrete version of the Cox proportional hazards model. For other analyses of rate data, see Breslow and Day (1987, Sec. 4.5), Frome (1983), and Hoem (1987). Doksum and Gasko (1990) summarized the connection between the logistic regression model and the proportional odds model for the analysis of survival data. Other articles dealing with loglinear and logistic models for grouped survival data include Aitkin and Clayton (1980), Aranda-Ordaz (1983), Larson (1984), Prentice and Gloeckler (1978), Schluchter and Jackson (1989), Stokes et al. (2012), and Thompson (1977).

Table 4.10 Data on Number of Deaths from Lung Cancer for Exercise

Follow-up Time Interval (months)	Disease Stage:	Histology ^a								
		I			II			III		
		1	2	3	1	2	3	1	2	3
0–2		9 (157)	12 (134)	42 (212)	5 77	4 71	28 130	1 21	1 22	19 101)
2–4		2 (139)	7 (110)	26 (136)	2 68	3 63	19 72	1 17	1 18	11 63)
4–6		9 (126)	5 (96)	12 (90)	3 63	5 58	10 42	1 14	3 14	7 43)
6–8		10 (102)	10 (86)	10 (64)	2 55	4 42	5 21	1 12	1 10	6 32)
8–10		1 (88)	4 (66)	5 (47)	2 50	2 35	0 14	0 10	0 8	3 21)
10–12		3 (82)	3 (59)	4 (39)	2 45	1 32	3 13	1 8	0 8	3 14)
12+		1 (76)	4 (51)	1 (29)	2 42	4 28	2 7	0 6	2 6	3 10)

^aValues in parentheses represent total follow-up time at risk.

Source: Reprinted from Holford (1980) with permission from the Biometric Society.

Section 4.5: Inference and Model Checking for Generalized Linear Models

4.4 Pearson statistics: For use of the Pearson statistic and related statistics for model comparison, see Agresti and Ryu (2010), Haberman (1977a), Lovison (2005), Pregibon (1982), Rao (1961), and Smyth (2003).

4.5 Diagnostics: McCullagh and Nelder (1989, Chap. 12) discussed model checking for GLMs. Separate diagnostics are useful for checking the adequacy of each component. For a family $g(\mu; \gamma)$ of link functions indexed by parameter γ , Pregibon (1980) showed how to estimate γ giving

the link with best fit and how to check the adequacy of a given link $g(\mu; \gamma_0)$. For discussions about residuals, see also Davison (1991), Green (1984), Pierce and Schafer (1986), Pregibon (1980, 1981), and Williams (1987). Pregibon (1982) showed that the squared standardized residual is the score statistic for testing whether the observation is an outlier. Davison and Hinkley (1997, Sec. 7.2) discussed bootstrapping in GLMs.

Section 4.6: Fitting Generalized Linear Models

4.6 IRLS/ML: Fisher (1935b) introduced the Fisher scoring method to calculate ML estimates for probit models. For further discussion of GLM model fitting and the relationship between iterative reweighted least squares and ML estimation, see Green (1984), Jørgensen (1983), McCullagh and Nelder (1989), and Nelder and Wedderburn (1972). Green (1984), Jørgensen (1983), and Palmgren and Ekholm (1987) also discussed this relation for exponential family nonlinear models.

Section 4.7: Quasi-likelihood and Generalized Linear Models

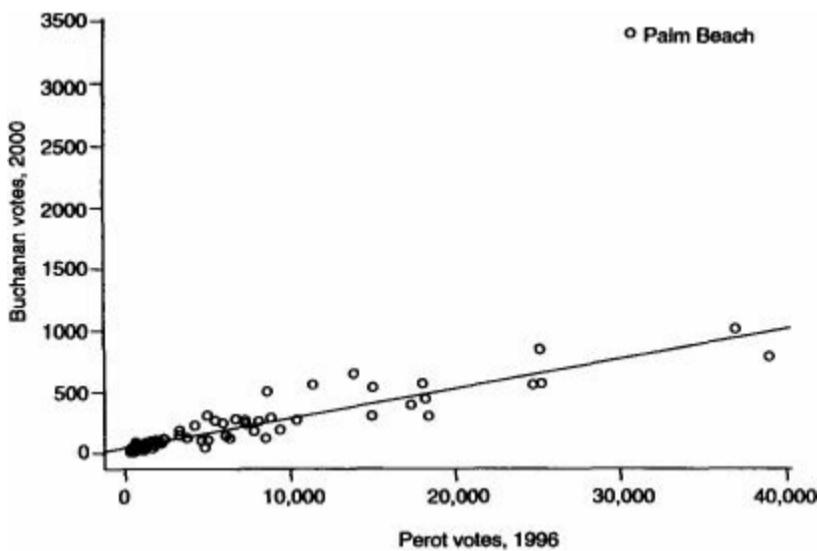
4.7 Quasi-likelihood: For more on quasi-likelihood, see Sections 12.3, 13.6.4, and 14.3, Breslow (1984, 1990), Cox (1983), Firth (1987), Hinde and Demétrio (1998), Heyde (1997), Lee et al. (2006, Chap. 3), McCullagh (1983), McCullagh and Nelder (1989), Nelder and Pregibon (1987), and Wedderburn (1974, 1976).

EXERCISES

Applications

4.1 In the 2000 U.S. presidential election, Palm Beach County in Florida was the focus of unusual voting patterns apparently caused by a confusing “butterfly ballot.” Many voters claimed that they voted mistakenly for the Reform Party candidate, Pat Buchanan, when they intended to vote for Al Gore. [Figure 4.7](#) shows the total number of votes for Buchanan plotted against the number of votes for the Reform Party candidate in 1996 (Ross Perot), by county in Florida.

[Figure 4.7](#) Total vote, by county in Florida, for Reform Party candidates Buchanan in 2000 and Perot in 1996.



a. In county i , let π_i denote the proportion of the vote for Buchanan and let x_i denote the proportion of the vote for Perot in 1996. For the linear probability model fitted to all counties except Palm Beach, $\hat{\mu}_i = -0.0003 + 0.0304x_i$. Give the value of P in the interpretation: The estimated proportion vote for Buchanan in 2000 was roughly $P\%$ of that for Perot in 1996.

b. For Palm Beach County, $\pi_i = 0.0079$ and $x_i = 0.0774$. Does this result appear to be an outlier? Investigate, by finding $\pi_i/\hat{\pi}_i$. (George W. Bush won the state by 537 votes and, with it, the Electoral College and the election. Other ballot design problems played a role in 110,000 disqualified “overvote” ballots, in which people mistakenly voted for more than one candidate, with Gore marked on 84,197 ballots and Bush on 37,731. For details, see A. Agresti and B. Presnell, *Statist. Sci.*, **17**: 436–440, 2003.)

4.2 For [Table 3.8](#) with scores (0, 0.5, 1.5, 4.0, 7.0) for alcohol consumption, ML fitting of the linear probability model for malformation has output:

Parameter	Estimate	Std Error	Wald 95% Conf Limits	
Intercept	0.0025	0.0003	0.0019	0.0032
alcohol	0.001087	0.000727	-0.0003	0.0025

a. Interpret the model parameter estimates. Use the fit to estimate the relative risk of malformation that compares alcohol consumption levels 0 and 7.0.

b. Some software (such as R) also reports $\hat{\beta} = 0.001087$ but instead reports $SE = 0.000832$. Why do you think the SE value is different? [Hint: The software with output shown above inverts the *observed* information matrix to obtain the standard errors.]

4.3 For [Table 4.2](#), refit the linear probability model or the logistic regression model using the scores (a) (0, 2, 4, 6), (b) (0, 1, 2, 3), and (c) (1, 2, 3, 4). Compare $\hat{\beta}$ for the three choices. Compare fitted values. Summarize the effect of linear transformations of scores, which preserve relative sizes of spacings between scores.

4.4 For the data shown in part in [Table 4.3](#), let $Y=1$ if a crab has at least one satellite, and $Y=0$ otherwise. Using $x = \text{weight}$, fit the linear probability model.

a. Use ordinary least squares. Interpret the parameter estimates. Find the estimated probability at the highest observed weight, 5.20 kg. Comment.

- b.** Try to fit the model using ML, treating Y as binomial. [The failure is due to a fitted probability falling outside the (0, 1) range. The fit in part (a) is ML for a *normal* random component, for which fitted values outside this range are permissible.]
- c.** Fit the logistic regression model. Show that the fitted probability at a weight of 5.20 kg equals 0.9968.

4.5 We use the following artificial data to illustrate comments in Section 4.5.3 about grouped versus ungrouped binary data:

X	Number of trials	Number of successes
0	4	1
1	4	2
2	4	4

Denote by M_0 the model $\text{logit}[P(Y = 1)] = \alpha$ and by M_1 the model $\text{logit}[P(Y = 1)] = \alpha + \beta x$. Denote the maximized log-likelihood values by L_0 for M_0 , L_1 for M_1 , and L_s for the saturated model. Create a data file in two ways, entering the data as (i) ungrouped data: $n_i = 1$, $i = 1, \dots, 12$; and (ii) grouped data: $n_i = 4$, $i = 1, 2, 3$.

- a.** Fit M_0 and M_1 for each data file. Report L_0 and L_1 in each case. Note they are the same for each form of data entry.
- b.** Show that the deviances for M_0 and M_1 differ for the two forms of data entry. Why is this? [Hint: How many parameters are in the saturated model for each data file?]
- c.** Show that the difference between the deviances for M_0 and M_1 is the same for each form of data entry. Why is this? (Thus, for testing the effect of x , it does not matter how you enter the data, but it does matter if you want to test goodness of fit.)

4.6 For [Table 4.3](#), [Table 4.8](#) shows SAS output for a Poisson loglinear model fit using x = weight and Y = number of satellites.

- a.** Use β to describe the weight effect. Show how to construct the reported confidence interval.
- b.** Construct a Wald test that Y is independent of x . Interpret. What else do you need to conduct a likelihood-ratio test of this hypothesis?
- c.** Is there evidence of overdispersion? Explain.

Table 4.8 SAS Output for Exercise 4.6

Criterion	DF	Value
Deviance	171	560.8664
Pearson Chi-Square	171	535.8957
Log Likelihood		71.9524

Parameter	Estimate	Std Error	Wald 95%	Conf Limits	Chi-Sq	Pr > ChiSq
Intercept	-0.4284	0.1789	-0.7791	-0.0777	5.73	0.0167
weight	0.5893	0.0650	0.4619	0.7167	82.15	<.0001

4.7 An experiment analyzes imperfection rates for two processes used to fabricate silicon wafers for computer chips. For treatment A applied to 10 wafers, the numbers of imperfections are 8, 7, 6, 6, 3, 4, 7, 2, 3, 4. Treatment B applied to 10 other wafers has 9, 9, 8, 14, 8, 13, 11, 5, 7, 6 imperfections. Treat the counts as independent Poisson variates having means μ_A and μ_B .

- a.** Fit the model $\log \mu = \alpha + \beta x$, where $x = 1$ for treatment B and $x = 0$ for treatment A. Show that $\exp(\beta) = \mu_B/\mu_A$ and interpret its estimate.
- b.** Test $H_0: \mu_A = \mu_B$ using the Wald or likelihood-ratio test of $H_0: \beta = 0$. Interpret.
- c.** Construct a 95% confidence interval for μ_B/μ_A . [Hint: First construct one for β .]
- d.** Test $H_0: \mu_A = \mu_B$ based on this result: If Y_1 and Y_2 are independent Poisson with means μ_1 and μ_2 , then given $n = (Y_1 + Y_2)$, Y_1 is $\text{bin}(n, \pi)$ with $\pi = \mu_1/(\mu_1 + \mu_2)$.

4.8 Refer to Exercise 4.7. The sample mean and variance are 5.0 and 4.2 for treatment A and 9.0 and 8.4 for treatment B.

- a.** Is there evidence of overdispersion for the Poisson model? Explain. Fit the negative

binomial loglinear model. Note that the estimated dispersion parameter is 0 and that estimates of treatment means and standard errors are the same as with the Poisson loglinear GLM.

b. For the overall sample of 20 observations, the sample mean and variance are 7.0 and 10.2. Fit the loglinear model having only an intercept term under Poisson and negative binomial assumptions. Compare results, and compare confidence intervals for the overall mean response. Why do they differ? [Note: This shows how the Poisson model can lose validity when an important covariate is unobserved.]

4.9 For the negative binomial model fitted to the crab satellite counts with log link and width predictor, $\hat{\beta} = -4.05$, $\hat{\beta} = 0.192$ ($SE = 0.048$), $\hat{\gamma} = 1.106$ ($SE = 0.197$) Interpret. Why is SE for $\hat{\beta}$ so different from $SE = 0.020$ for the corresponding Poisson GLM in Section 4.3.2? Which is more appropriate? Why?

4.10 Fit the rate model (4.15) for the heart valve operations.

- a.** Find the 95% profile likelihood confidence interval for β_1 , and show that it translates to (1.32, 10.39) for the true multiplicative effect $\exp(\beta_1)$.
- b.** Show that the deviance comparing $\{y_{ij}\}$ with $\{\hat{\mu}_{ij}\}$ is $G^2 = 3.22$, with residual df = 1. Show that $G^2 = 1.09$ for the corresponding model with identity link.
- c.** What is the effect on the model parameter estimates, SE values, and the deviance when (i) the times at risk are doubled, but the numbers of deaths stay the same; (ii) the times at risk stay the same, but the numbers of deaths double; and (iii) the times at risk and the numbers of deaths both double?

4.11 Table 4.9 is based on a study with British doctors.

- a.** For each age, find the sample coronary death rates per 1000 person-years for nonsmokers and smokers. To compare them, take their ratio and describe its dependence on age.
- b.** Fit a main-effects model for the log rates using age and smoking as factors. In discussing lack of fit, show that this model assumes a constant ratio of nonsmokers' to smokers' coronary death rates over age.
- c.** From part (a), explain why it is sensible to add a quantitative interaction of age and smoking. For this model, show that the log ratio of coronary death rates changes linearly with age. Assign scores to age, fit the model, and interpret.

Table 4.9 Data for Exercise 4.11 on Coronary Death Rates

Age	Person-Years		Coronary Deaths	
	Nonsmokers	Smokers	Nonsmokers	Smokers
35–44	18,793	52,407	2	32
45–54	10,673	43,248	12	104
55–64	5710	28,612	28	206
65–74	2585	12,663	28	186
75–84	1462	5317	31	102

Source: R. Doll and A. Bradford Hill, *Natl. Cancer Inst. Monogr.* **19**: 205–268, 1966. See also N. R. Breslow in *A Celebration of Statistics*, ed. A. C. Atkinson and S. E. Fienberg. New York: Springer-Verlag, 1985.

4.12 Table 4.10 describes survival for 539 males diagnosed with lung cancer. The prognostic factors are histology (H) and state (S) of disease. The assumption of a constant rate over time is often not sensible, and this study divided the time scale (T) into two-month intervals and let the rate vary by the time interval. Let μ_{ijk} denote the expected number of deaths and t_{ijk} the total time at risk for histology i and state of disease j , in follow-up time interval k . Analyses suggested a lack of interaction between T and either prognostic factor (i.e., such *proportional hazards* models have the same effects of H and S for each time interval).

a. The main effects model

$$\log(\mu_{ijk}/t_{ijk}) = \alpha + \beta_i^H + \beta_j^S + \beta_k^T$$

has deviance 43.9. Explain why df = 52. Does the model seem to fit adequately?

b. For this model, interpret the estimated effects of $\hat{\beta}_2^S - \hat{\beta}_1^S = 0.470$ ($SE = 0.174$), $\hat{\beta}_3^S - \hat{\beta}_1^S = 1.324$ ($SE = 0.152$).

c. The model that adds an $S \times H$ interaction term has deviance 41.5 with $df = 48$. Test whether a significantly improved fit results by allowing this interaction.

4.13 [Table 4.11](#) shows the three-point shooting, by game, of Ray Allen of the Boston Celtics during the 2010 NBA (basketball) playoffs. Commentators remarked that his shooting varied dramatically from game to game. In game i , suppose that Y_i = number of three-point shots made out of n_i attempts is a $\text{bin}(n_i, \pi_i)$ variate and the $\{Y_i\}$ are independent.

Table 4.11 Data for Exercise 4.13 on Basketball Shooting

Game	Number Made	Number of Attempts	Game	Number Made	Number of Attempts	Game	Number Made	Number of Attempts
1	0	4	9	1	8	17	3	7
2	7	9	10	6	9	18	0	2
3	4	11	11	0	5	19	8	11
4	3	6	12	2	5	20	0	8
5	5	6	13	0	5	21	0	4
6	2	7	14	2	4	22	0	4
7	3	7	15	5	7	23	2	5
8	0	1	16	1	3	24	2	7

Source: boston.stats.com/nba.

a. Fit the model, $\pi_i = \alpha$, and find and interpret $\hat{\alpha}$ and its standard error. Does the model appear to fit adequately? [Note: You could check this with a small-sample test of independence of the 24×2 table of game and the binomial outcome.]

b. Adjust the standard error for overdispersion. Using the original SE and its correction, find and compare 95% confidence intervals for α . Interpret

c. Describe a factor that could cause overdispersion. [Hint: Is it realistic to treat the success probability as identical from shot to shot?]

4.14 Refer to [Table 14.6](#). Fit a loglinear model with an indicator variable for race, (a) assuming a Poisson distribution, and (b) allowing overdispersion with a quasi-likelihood approach. Compare results.

Theory and Methods

4.15 Describe the purpose of the link function of a GLM. Explain why the identity link is not often used with binomial or Poisson responses.

4.16 For binary data, define a GLM using the log link. Show that effects refer to the relative risk. Why do you think this link is not often used? [Hint: What happens if the linear predictor takes a positive value?]

4.17 For the logistic regression model [\(4.6\)](#) with $\beta > 0$, show that (a) as $x \rightarrow \infty$, $\pi(x)$ is monotone increasing, and (b) the curve for $\pi(x)$ is the cdf of a logistic distribution having mean $-\alpha/\beta$ and standard deviation $\pi/(|\beta|\sqrt{3})$.

4.18 Let Y_i be a $\text{bin}(n_i, \pi_i)$ variate for group $i, i = 1, \dots, N$, with $\{Y_i\}$ independent. For the model that $\pi_1 = \dots = \pi_N$, denote that common value by π . For observations $\{y_i\}$, show that $\hat{\pi} = (\sum_i y_i) / (\sum_i n_i)$. When all $n_i = 1$, for testing this model's fit in the $N \times 2$ table, show that $X^2 = N$. Thus, goodness-of-fit statistics can be completely uninformative for ungrouped data. (See also Exercise 5.35.)

4.19 Suppose that Y_i is Poisson with $g(\mu_i) = \alpha + \beta x_i$, where $x_i = 1$ for $i = 1, \dots, n_A$ from group A and $x_i = 0$ for $i = n_A + 1, \dots, n_A + n_B$ from group B. Show that for any link function g , the GLM likelihood [equations \(4.25\)](#) imply that fitted means $\hat{\mu}_A$ and $\hat{\mu}_B$ equal the sample means.

4.20 A method for negative exponential modeling of survival times relates to the Poisson loglinear model for rates (Aitkin and Clayton 1980). Let T denote the time to some event, with pdf f and cdf F . For subject i , let $w_i = 1$ for death and 0 for censoring, and let $T = \sum_i t_i$, and $W = \sum_i w_i$.

a. Explain why the survival-time log likelihood for n independent observations is

$$L(\lambda) = \sum_i w_i \log[f(t_i)] + \sum_i (1 - w_i) \log[1 - F(t_i)].$$

(This actually applies only for *noninformative* censoring mechanisms.) Assuming $f(t) = \lambda \exp(-\lambda t)$, show that $\hat{\lambda} = W/T$. Conditional on T , explain why W has a Poisson distribution with mean $T\lambda$, and using the Poisson likelihood show that $\hat{\lambda} = W/T$.

b. The *hazard function* represents the instantaneous rate of death for subjects who have survived to time t . Suppose that $h(t; \mathbf{x}) = \lambda \exp(\boldsymbol{\beta}^T \mathbf{x})$. With parameter λ in $f(t)$ replaced by $\lambda \exp(\boldsymbol{\beta}^T \mathbf{x})$ and with $\mu_i = t_i \lambda \exp(\boldsymbol{\beta}^T \mathbf{x}_i)$, show that L simplifies to

$$L(\lambda, \boldsymbol{\beta}) = \sum_i w_i \log \mu_i - \sum_i \mu_i - \sum_i w_i \log t_i.$$

c. Explain why maximizing $L(\lambda, \boldsymbol{\beta})$ is equivalent to maximizing the likelihood for the Poisson loglinear model

$$\log \mu_i - \log t_i = \log \lambda + \boldsymbol{\beta}^T \mathbf{x}_i$$

with offset $\log(t_i)$, using “observations” $\{w_i\}$.

d. When we sum terms in L for subjects having a common value of x , explain why the observed data are the numbers of deaths ($\sum_i w_i$) at each setting of x , and the offset is the log of $(\sum_i t_i)$ at each setting.

4.21 A binomial GLM $\pi_i = \Phi(\sum_j \beta_j x_{ij})$ with arbitrary inverse link function Φ assumes that $n_i Y_i$ has a $\text{bin}(n_i, \pi_i)$ distribution. Find w_i in (4.29) and hence $\text{cov}(\boldsymbol{\beta})$. For logistic regression, show that $w_i = n_i \pi_i (1 - \pi_i)$.

4.22 For the class of binary models (4.8) and (4.9), suppose the standard cdf Φ corresponds to a pdf φ that is symmetric around 0.

a. Show that x at which $\pi(x) = 0.50$ is $x = -\alpha/\beta$.

b. Show that the rate of change in $\pi(x)$ when $\pi(x) = 0.50$ is $\beta\varphi(0)$. Show this is 0.25β for the logit link and $\beta/\sqrt{2\pi}$ (where $\pi = 3.14\dots$) for the probit link.

c. Show that the probit regression curve for $\beta > 0$ has the shape of a normal cdf with mean $-\alpha/\beta$ and standard deviation $1/|\beta|$.

4.23 For binary observations, consider the model $\pi(x) = \frac{1}{2} + (1/\pi) \tan^{-1}(\alpha + \beta x)$.

a. Show that this corresponds to a cdf of a distribution for which the standard version is the Cauchy. When would you expect a GLM using this curve to be more appropriate than logistic regression?

b. Explain how this model generalizes to a family of models for which the link function is the inverse of the cdf of a t distribution for some df value, the probit resulting as $\text{df} \rightarrow \infty$.

4.24 A GLM has parameter β with sufficient statistic S . A goodness-of-fit test statistic T has observed value t_o . If β were known, a P -value is $P = P(T \geq t_o; \beta)$. Explain why $P(T \geq t_o | S)$ is the uniform minimum variance unbiased estimator of P .

4.25 Find the form of the deviance residual (4.42) for an observation in a **(a)** binomial GLM, and **(b)** Poisson GLM. Illustrate part (b) for a cell count in a two-way contingency table for the model of independence.

4.26 Let y_{ij} be observation j of a count variable for group $i, i = 1, \dots, I, j = 1, \dots, n_i$. Suppose that $\{Y_{ij}\}$ are independent Poisson with $E(Y_{ij}) = \mu_i$.

a. Show that the ML estimate of μ_i is $\hat{\mu}_i = \bar{y}_i = \sum_j y_{ij}/n_i$.

b. Simplify the expression for the deviance for this model. [For testing this model, it follows from Fisher (1970, p. 58, originally published in 1925) that the deviance and the Pearson statistic $\sum_i \sum_j (y_{ij} - \bar{y}_i)^2 / \bar{y}_i$ have approximate chi-squared distributions with $\text{df} = \sum_i (n_i - 1)$. For a single group, Cochran (1954) referred to $\sum_j (y_{1j} - \bar{y}_1)^2 / \bar{y}_1$ as the *variance test* for the fit of a Poisson distribution, since it compares the sample variance to the estimated Poisson variance \bar{y}_1 .]

4.27 For known k , show that the negative binomial distribution (4.13) has exponential family form (4.1) with natural parameter $\log[\mu/(\mu + k)]$.

4.28 Consider the normal distribution $N(\mu, \sigma^2)$

- a. With fixed σ , show it satisfies exponential family (4.1), and identify the components. Formulate the ordinary regression model as a GLM. Explain why the least-squares estimates are then ML and the deviance is $\sum_i(y_i - \hat{\mu}_i)^2$.

- b. When σ is also a parameter, show that it satisfies the exponential dispersion family (4.17).

4.29 For a GLM, refer to the adjusted response variable in Section 4.6.4. Let $z_0 = \hat{W}^{1/2}\hat{z}$, $X_0 = \hat{W}^{1/2}X$ and $\hat{\eta}_0 = \hat{W}^{1/2}X\hat{\beta}$. Show that (a) $\hat{\beta}$ is the ordinary least-squares solution for the model $z_0 = X_0\beta + \epsilon$, (b) the estimated hat matrix for the GLM equals $X_0(X_0^T X_0)^{-1}X_0^T$, (c) $(z_0 - \hat{\eta}_0)$ are the Pearson residuals, and (d) z_0 and η_0 are analogs for GLMs of y and $\hat{\mu}$ in ordinary linear models, in terms of projections and orthogonal decompositions. (Thanks to Gianfranco Lovison for suggesting these results. See also Fahrmeir and Tutz 2001, pp. 147–148, and Tutz 2011, Sec. 3.10.)

4.30 Let $\beta^{(0)}$ denote an initial guess for the value $\hat{\beta}$ that maximizes a function $L(\beta)$.

- a. Using $L'(\hat{\beta}) = L'(\beta^{(0)}) + (\hat{\beta} - \beta^{(0)})L''(\beta^{(0)}) + \dots$, argue that for $\beta^{(0)}$ close to $\hat{\beta}$, approximately $0 = L'(\beta^{(0)}) + (\hat{\beta} - \beta^{(0)})L''(\beta^{(0)})$. Solve this equation to obtain an approximation $\beta^{(1)}$ for $\hat{\beta}$.
- b. Let $\beta^{(t)}$ denote approximation t for $\hat{\beta}$, $t = 0, 1, 2, \dots$. Justify that the next approximation is $\beta^{(t+1)} = \beta^{(t)} - L'(\beta^{(t)})/L''(\beta^{(t)})$.

4.31 For n independent observations from a Poisson distribution, show that Fisher scoring gives $\mu^{(t+1)} = \bar{y}$ for all $t > 0$. By contrast, what happens with Newton–Raphson?

4.32 Write a computer program using the Newton–Raphson algorithm to maximize the likelihood for a binomial sample. For $\pi = 0.3$ based on $n = 10$, print out results of the first six iterations when the starting value $\pi^{(0)}$ is (a) 0.1, (b) 0.2, ..., 0.9. Summarize the effects of the starting value on speed of convergence. What happens if π is 0 or 1?

4.33 For noncanonical links in a GLM, show that the observed information matrix depends on the data and hence differs from the expected information. Illustrate using the probit model.

4.34 Suppose $y_{i1}, y_{i2}, \dots, y_{in_i}$ are responses on a binary survey question for n_i members of a family, with $P(Y_{ij} = 1) = \pi = 1 - P(Y_{ij} = 0)$, $j = 1, \dots, n_i$, $i = 1, \dots, N$.

- a. Overdispersion: In each family, suppose everyone sees the response by the “head of household” and then makes the same response. State the distribution of $\sum_j Y_{ij}$ and compare its variance to that of the binomial.
- b. Underdispersion: In each family, suppose member j hears the response of member $j - 1$ and makes the opposite response, $j = 2, \dots, n_i$, with n_i , an even number. State the distribution of $\sum_j Y_{ij}$ and compare its variance to that of the binomial.

4.35 Sometimes, sample proportions are continuous rather than of the binomial form (number of successes)/(number of trials). Each observation is any real number between 0 and 1, such as the proportion of a tooth surface that is covered with plaque. For independent responses $\{y_i\}$, Aitchison and Shen (1980), Bartlett (1937), and Lesaffre et al. (2007) modeled $\text{logit}(Y_i) \sim N(\beta_i, \sigma^2)$. Then Y_i itself has a *logit-normal distribution* (Section 1.6.2).

- a. Expressing a $N(\beta, \sigma^2)$ variate as $\beta + \sigma Z$, where Z is standard normal, show that $Y_i = \exp(\beta_i + \sigma Z)/[1 + \exp(\beta_i + \sigma Z)]$.

- b. Show that for small σ ,

$$Y_i = \frac{e^{\beta_i}}{1 + e^{\beta_i}} + \frac{e^{\beta_i}}{1 + e^{\beta_i}} \frac{1}{1 + e^{\beta_i}} \sigma Z + \frac{e^{\beta_i}(1 - e^{\beta_i})}{2(1 + e^{\beta_i})^3} \sigma^2 Z^2 + \dots$$

- c. Letting $\mu_i = e^{\beta_i}/(1 + e^{\beta_i})$, when σ is close to 0 show that

$$E(Y_i) \approx \mu_i, \quad \text{var}(Y_i) \approx [\mu_i(1 - \mu_i)]^2 \sigma^2.$$

- d. For independent continuous proportions $\{y_i\}$, let $\mu_i = E(Y_i)$. For a GLM, it is sensible to use an inverse cdf link for μ_i , but it is unclear how to choose a distribution for Y_i . The approximate moments for the logit-normal motivate a quasi-likelihood approach (Wedderburn 1974) with variance function $v(\mu_i) = \phi[\mu_i(1 - \mu_i)]^2$ for unknown ϕ . Explain why this provides

similar results as fitting a normal regression model to the sample logits assuming constant variance. (The QL approach has the advantage of not requiring adjustment of 0 or 1 observations, for which sample logits do not exist.)

e. Wedderburn (1974) modeled data on the proportion of a leaf showing a type of blotch. Envision an approximation of binomial form based on cutting each leaf into a large number of small regions of the same size and observing for each region whether it is mostly covered with blotch. Explain why this suggests that $v(\mu_i) = \phi\mu_i(1 - \mu_i)$. What violation of the binomial assumptions might make this questionable? [The parametric family of beta distributions has variance function of this form (see Section 14.3.1 and Cox 1996).]

¹Its kurtosis equals that of a t distribution with $df = 9$. Albert and Chib (1993) noted that a t variate with $df = 8$ divided by 0.634 well approximates a standard logistic variate. Caffo and Griswold (2006) used a similar approximation with $df = 8.78$.

²See en.wikipedia.org/wiki/Horseshoe_crab and horseshoecrab.org for details about horseshoe crabs, including pictures of their mating.

³This result, which follows from the simple form of the likelihood equations for such models, is shown for Poisson loglinear models in Section 10.2.3.

⁴Thanks to Dr. Gianfranco Lovison for showing me this argument.

CHAPTER 5

Logistic Regression

In introducing generalized linear models for binary data in Chapter 4 we highlighted logistic regression. This is the most important model for categorical response data, being commonly used for a wide variety of applications.

Early uses of logistic regression were in biomedical studies, for instance, to model whether subjects have a particular condition such as lung cancer. The past 25 years have seen much use in social science research, for modeling opinions and behavior decisions, and in business applications. In *credit-scoring*, logistic regression is used to model the probability that a subject is credit worthy. For instance, the probability that a subject pays a bill on time may use predictors such as the size of the bill, annual income, occupation, mortgage and debt obligations, percentage of bills paid on time in the past, and other aspects of an applicant's credit history. Another area of increasing application is genetics, such as to estimate quantitative trait loci effects by modeling the probability that an offspring inherits an allele of one type instead of another type as a function of phenotypic values on various traits for that offspring.

In this chapter we study logistic regression more closely. Section 5.1 discusses parameter interpretation. In Section 5.2 we present inferential methods for those parameters. Sections 5.3 and 5.4 generalize the model to multiple predictors, which may be quantitative and/or qualitative. Finally, in Section 5.5 we apply GLM fitting methods to specify and solve likelihood equations for logistic regression.

5.1 INTERPRETING PARAMETERS IN LOGISTIC REGRESSION

For a binary response variable Y and an explanatory variable X , let $\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$. The logistic regression model is

$$(5.1) \quad \pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}.$$

Equivalently, the *logit* (log odds) has the linear relationship

$$(5.2) \quad \text{logit}[\pi(x)] = \log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x.$$

5.1.1 Interpreting β : Odds, Probabilities, and Linear Approximations

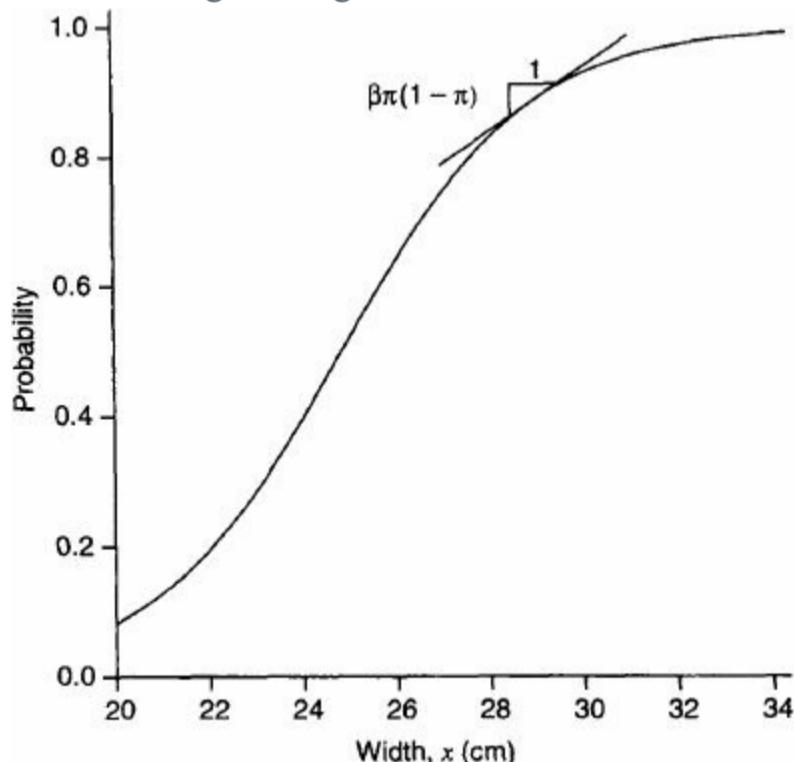
How can we interpret β in (5.2)? Its sign determines whether $\pi(x)$ is increasing or decreasing as x increases. The rate of climb or descent increases as $|\beta|$ increases; as $\beta \rightarrow 0$ the curve flattens to a horizontal straight line. When $\beta = 0$, Y is independent of X . For quantitative x with $\beta > 0$, the curve for $\beta(x)$ has the shape of the cdf of the logistic distribution (recall Section 4.2.5). Since the logistic density is symmetric, $\pi(x)$ approaches 1 at the same rate that it approaches 0.

Exponentiating both sides of (5.2) shows that the odds are an exponential function of x . This provides a basic interpretation for the magnitude of β : The odds multiply by e^β for every 1-unit increase in x . In other words, e^β is an odds ratio, the odds at $X = x + 1$ divided by the odds at $X = x$.

Many scientists are not familiar with odds or logits, so the interpretation of a multiplicative effect of e^β on the odds scale or an additive effect of β on the logit scale is not helpful to them. A simpler slope interpretation uses a linearization argument (Berkson 1951). Since it has a curved rather than a linear appearance, the logistic regression function (5.1) implies that the rate of change in $\pi(x)$ per unit change in x varies. A straight line drawn tangent to the curve at a particular x value, shown in Figure 5.1, describes the instantaneous rate of change at that point. Calculating $\frac{\partial \pi(x)}{\partial x}$ with (5.1) yields a fairly complex function of the parameters and x , but it simplifies to the form $\beta\pi(x)[1 - \pi(x)]$.

For instance, the line tangent to the curve at x for which $\pi(x) = \frac{1}{2}$ has slope $\beta(\frac{1}{2})(\frac{1}{2}) = \beta/4$; when $\pi(x) = 0.9$ or 0.1, it has slope 0.09β . The slope approaches 0 as $\pi(x)$ approaches 1.0 or 0. The steepest slope occurs at x for which $\pi(x) = \frac{1}{2}$; that x value is $x = -\alpha/\beta$. [To check that $\pi(x) = \frac{1}{2}$ at this point, substitute $-\alpha/\beta$ for x in (5.1), or substitute $\pi(x) = \frac{1}{2}$ in (5.2) and solve for x .] This x value is sometimes called the *median effective level*. In toxicology studies it is called LD₅₀ (LD = lethal dose), the dose with a 50% chance of a lethal result.

Figure 5.1 Linear approximation to logistic regression curve.



From this linear approximation, near x where $\pi(x) = \frac{1}{2}$, a change in x of $1/\beta$ corresponds to a change in $\pi(x)$ of roughly $(1/\beta)(\beta/4) = \frac{1}{4}$, that is, $1/\beta$ approximates the distance between x values where $\pi(x) = 0.50$ and where $\pi(x) = 0.25$ or 0.75 (in reality, 0.27 and 0.73). The linear approximation works better for smaller changes in x , however. Since the rate of change varies according to the value of x , a summary of them is the average of $\beta\pi(x_i)[1 - \pi(x_i)]$ for the subjects in the sample.

An alternative way to interpret the effect reports the values of $\pi(x)$ at certain x values, such as at the minimum and maximum values. To do this, we substitute the values for x into formula (5.1) for $\pi(x)$. It is more resistant to outliers on x to report the $\pi(x)$ values at the quartiles of x than at the

extremes. The change in $\pi(x)$ over the middle half of x values, from the lower quartile to the upper quartile, is a useful summary of the effect. It can be compared with the corresponding change over the middle half of values of other quantitative predictors.

The intercept parameter α is not usually of particular interest. However, by centering the predictor about 0 [i.e., replacing x by $(x - \bar{x})$], α becomes the logit at $x = \bar{x}$, and thus $e^\alpha / (1 + e^\alpha) = \pi(\bar{x})$. As in ordinary regression, centering is also helpful in complex models containing quadratic or interaction terms to reduce correlations among model parameter estimates.

5.1.2 Looking at the Data

In practice, these interpretations use formula (5.1) with ML estimates substituted for parameters. Before fitting the model and making such interpretations, look at the data to check that the logistic regression model is appropriate. Since y takes only values 0 and 1, it is difficult to check this by an ordinary scatterplot of observed (x, y) values.

It can be helpful to plot sample proportions or logits against x . Let n_i denote the number of observations at setting i of x . Of them, let y_i denote the number of “1” outcomes, with $p_i = y_i/n_i$. Sample logit (also called *empirical logit*) i is $\log[p_i/(1 - p_i)] = \log[y_i/(n_i - y_i)]$. The scatterplot of sample logits should be roughly linear. The sample logit is not finite when $y_i = 0$ or n_i . An ad hoc adjustment adds a positive constant to the number of outcomes of the two types. The adjustment

$$\log \frac{y_i + \frac{1}{2}}{n_i - y_i + \frac{1}{2}}$$

is the least-biased estimator of this form for the true logit (see Note 5.2).

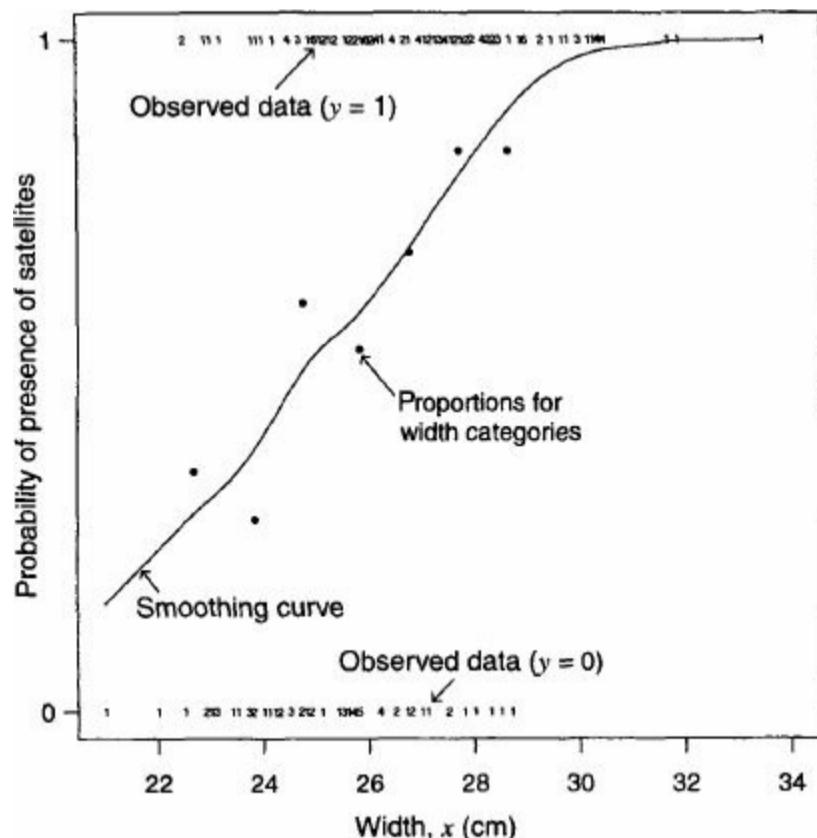
When x is continuous and all $n_i = 1$, or when x is essentially continuous and all n_i are small, this is unsatisfactory. We could group the data with nearby x values into categories before calculating sample proportions and sample logits. A better approach that does not require choosing arbitrary categories uses a smoothing mechanism to reveal trends. One such smoothing approach fits a *generalized additive model* (to be introduced in Section 7.4.9), which replaces the linear predictor of a GLM by a smooth function. A plot of this fit reveals whether severe discrepancies occur from the S-shaped trend predicted by logistic regression.

5.1.3 Example: Horseshoe Crab Mating Revisited

To illustrate logistic regression, we reanalyze the horseshoe crab data introduced in Section 4.3.2. The binary response is whether a female crab has any male crabs residing nearby (satellites). For crab i , let $y_i = 1$ if she has at least one satellite and $y_i = 0$ if she has none. Here, we use as a predictor the female crab's carapace width.

[Figure 5.2](#) plots the data against $x = \text{width}$. The scatterplot consists of a set of points with $y_i = 1$ and a second set of points with $y_i = 0$. The numbered symbols indicate the number of observations at each point. It appears that $y_i = 1$ tends to occur relatively more often at higher x values; in fact, all crabs with width > 29 cm have satellites. The positive effect of width is also suggested by the grouping of the data used to investigate adequacy of Poisson regression models in Section 4.3.3 ([Table 4.4](#)). In each of the eight width categories, we computed the sample proportion of crabs having satellites and the mean width for the crabs in that category. [Figure 5.2](#) shows eight dots representing the sample proportions of female crabs having satellites plotted against the mean widths for the eight categories. [Figure 5.2](#) also shows a curve based on smoothing the data using the generalized additive modeling method, assuming a binomial response and logit link. This curve shows a roughly increasing trend and is more informative than viewing the binary data alone. It suggests that an S-shaped regression function may describe this relationship relatively well. Since the eight plotted sample proportions and the GAM smoothing curve both suggest an increasing trend, we proceed with fitting the logistic regression model with linear width predictor.

[Figure 5.2](#) Whether satellites are present (1 = yes, 0 = no) by width of female crab, with smoothing fit of generalized additive model.



We defer to Section 5.5 the details about ML fitting. Software (see the text website) reports output such as [Table 5.1](#) exhibits. Let $\pi(x)$ denote the probability that a female horseshoe crab of width x has a satellite. The ML fit is

[Table 5.1](#) Output (Based on SAS) for Logistic Regression Model with Horseshoe Crab Data

Criteria For Assessing Goodness Of Fit					
Criterion	DF	Value			
Deviance	171	194.4527			
Pearson Chi-Square	171	165.1434			
Log Likelihood		-97.2263			

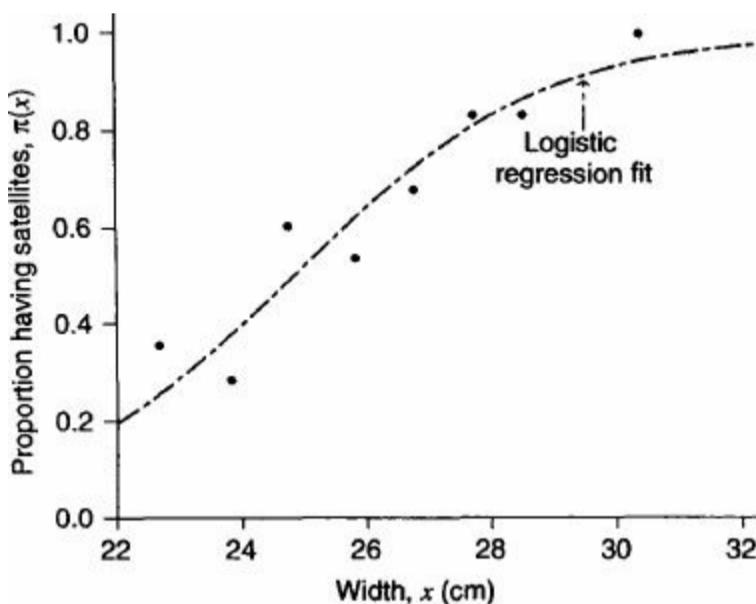
Parameter	Estimate	Std Error	Likelihood-Ratio		Wald	
			95% Conf Limits		Chi-Sq	P>ChiSq
Intercept	-12.3508	2.6287	-17.8097	-7.4573	22.07	<.0001
width	0.4972	0.1017	0.3084	0.7090	23.89	<.0001

$$\hat{\pi}(x) = \frac{\exp(-12.351 + 0.497x)}{1 + \exp(-12.351 + 0.497x)}.$$

Substituting $x = 26.3$ cm, the mean width level in this sample, $\hat{\pi}(x) = 0.674$. The estimated probability equals $\frac{1}{2}$ when $x = -\hat{\alpha}/\hat{\beta} = 12.351/0.497 = 24.8$.

[Figure 5.3](#) plots $\hat{\pi}(x)$ from the logistic fit against width, again superimposing the sample proportions that we viewed in [Figure 5.2](#). The curve seems to follow reasonably well the trend in those proportions.

[Figure 5.3](#) Logistic regression fitted curve and sample proportions of satellites, by width of female crab.



The estimated odds of a satellite multiply by $\exp(\hat{\beta}) = \exp(0.497) = 1.64$ for each 1-cm increase in width; that is, there is a 64% increase. To convey the effect less technically, we could report the incremental rate of change in the probability of a satellite. At the mean width, $\hat{\pi}(x) = 0.674$, and $\hat{\pi}(x)$ increases by about $\hat{\beta}[\hat{\pi}(x)(1 - \hat{\pi}(x))] = 0.497(0.674)(0.326) = 0.11$ for a 1-cm increase in width. Or, we could report $\hat{\pi}(x)$ at the quartiles of x . The lower quartile, median, and upper quartile for width are 24.9, 26.1, and 27.7; $\hat{\pi}(x)$ at those values equals 0.51, 0.65, and 0.81, increasing by 0.30 over the x values for the middle half of the sample.

The latter summary is useful for comparing the effects of predictors having different units. For instance, with the female crab's weight as the predictor, $\text{logit}[\hat{\pi}(x)] = -3.695 + 1.815x$. A 1-kg increase in weight is not comparable to a 1-cm increase in width, so $\hat{\beta} = 1.815$ for $x = \text{weight}$ is not comparable to $\hat{\beta} = 0.497$ for $x = \text{width}$. The quartiles for weight are 2.00, 2.35, and 2.85; $\hat{\pi}(x)$ at those values are 0.48, 0.64, and 0.81, increasing by 0.33 over the middle half of the sampled weights. The effect is similar to that of width, which is not surprising as these predictors are very highly correlated.

5.1.4 Logistic Regression with Retrospective Studies

Another property of logistic regression relates to situations in which the explanatory variable X rather than the response variable Y is random. This occurs with retrospective sampling designs, such as case-control biomedical studies. For samples of subjects having $y = 1$ (cases) and having $y = 0$ (controls), the value of X is observed. Evidence exists of an association if the distribution of X values differs between cases and controls. In retrospective studies, we can estimate odds ratios. Effects in the logistic regression model refer to odds ratios. We can fit such models and estimate effects in case-control studies (Breslow and Powers 1978, Prentice and Pyke 1979).

Here is a justification for this. Let Z indicate whether a subject is sampled ($1 = \text{yes}$, $0 = \text{no}$). Let $\rho_1 = P(Z = 1|y = 1)$ denote the probability of sampling a case, and let $\rho_0 = P(Z = 1|y = 0)$ denote the probability of sampling a control. Even though the conditional distribution of Y given $X = x$ is not sampled, we need a model for $P(Y = 1|z = 1, x)$, assuming that $P(Y = 1|x)$ follows the logistic model. By Bayes' theorem,

$$(5.3) \quad P(Y = 1|z = 1, x) = \frac{P(Z = 1|y = 1, x)P(Y = 1|x)}{\sum_{j=0}^1 [P(Z = 1|y = j, x)P(Y = j|x)]}.$$

Now, suppose that $P(Z = 1|y, x) = P(Z = 1|y)$ for $y = 0$ and 1 ; that is, for each y , the sampling probabilities do not depend on x . For instance, often x refers to exposure of some type, such as whether someone has been a smoker. Then, for cases and for controls, the probability of being sampled is the same for smokers and nonsmokers. Under this assumption, substituting ρ_1 and ρ_0 in (5.3) and dividing numerator and denominator by $P(Y = 0|x)$, we get

$$P(Y = 1|z = 1, x) = \frac{\rho_1 \exp(\alpha + \beta x)}{\rho_0 + \rho_1 \exp(\alpha + \beta x)}.$$

Then, dividing numerator and denominator by ρ_0 and using $\rho_1/\rho_0 = \exp[\log(\rho_1/\rho_0)]$ yields

$$\text{logit}[P(Y = 1|z = 1, x)] = \alpha^* + \beta x$$

with $\alpha^* = \alpha + \log(\rho_1/\rho_0)$. The logistic regression model holds with the same effect parameter β as in the model for $P(Y = 1|x)$. If the sampling rate for cases is greater than that for controls, the intercept estimated is larger than the one estimated with a prospective study.

With case-control studies, it is not possible to estimate β in binary-response models with links other than the logit. Unlike the odds ratio, the effect for the conditional distribution of X given y does not then equal that for Y given x . This is an important advantage of the logit link and is one reason why logistic regression models are so popular in biomedical studies.

Many case-control studies employ matching. Each case is matched with one or more control subjects. The controls are like the case on key characteristics such as age. The model and subsequent analysis should take the matching into account. In Section 11.2.5 we discuss logistic regression for matched case-control studies.

5.1.5 Logistic Regression Is Implied by Normal Explanatory Variables

Regardless of the sampling mechanism, logistic regression may or may not describe a relationship well. In one special case, it necessarily holds. Given that $Y = i$, suppose that X has $N(\mu_i, \sigma^2)$ distribution, $i = 0, 1$. Then, by Bayes' theorem, $P(Y = 1|X = x)$ satisfies the logistic model with $\beta = (\mu_1 - \mu_0)/\sigma^2$ (Cornfield 1962). Thus, when a population is a mixture of two types of subjects, one type with $y = 1$ that is approximately normally distributed on X and the other type with $y = 0$ that is approximately normal on X with similar variance, the logistic regression function approximates well the curve for $\pi(x)$.

The result extends to a vector of explanatory variables having multivariate normal distributions in each case (Exercise 5.30 and Section 15.1.1). If the distributions are normal but with different variances, the model applies but having a quadratic term. In that case, the relationship is nonmonotone, with $\pi(x)$ increasing and then decreasing, or the reverse.

5.2 INFERENCE FOR LOGISTIC REGRESSION

By standard results, ML estimators of logistic regression model parameters have large-sample normal distributions. Inference can use the (Wald, likelihood-ratio, score) triad of methods introduced in Section 1.3.3.

5.2.1 Inference About Model Parameters and Probabilities

For the logistic model with a single predictor,

$$\text{logit}[\pi(x)] = \alpha + \beta x,$$

significance tests focus on $H_0: \beta = 0$, the hypothesis of independence. The Wald test uses the log likelihood at $\hat{\beta}$, with test statistic $z = \hat{\beta}/SE$ or its square; under H_0 , z^2 is asymptotically χ^2_1 . The likelihood-ratio test uses twice the difference between the maximized log likelihood at $\hat{\beta}$ and at $\beta = 0$ and also has an asymptotic χ^2_1 null distribution. The score test uses the log likelihood at $\beta = 0$ through the derivative of the log likelihood (i.e., the score function) at that point. The test statistic compares the sufficient statistic for β to its null expected value, suitably standardized [$N(0,1)$ or χ^2_1]. In Section 5.3.5 we present this test.

A confidence interval for β results from inverting a test of $H_0: \beta = \beta_0$. The interval is the set of β_0 for which the chi-squared test statistic is no greater than $\chi^2_1(\alpha) = z_{\alpha/2}^2$. For the Wald approach, this means $[(\hat{\beta} - \beta_0)/SE]^2 \leq z_{\alpha/2}^2$, so the interval is $\hat{\beta} \pm z_{\alpha/2}(SE)$.

For summarizing the relationship, other characteristics may have greater importance than β , such as $\pi(x)$ at various x values. For fixed $x = x_0$, $\text{logit}[\hat{\pi}(x_0)] = \hat{\alpha} + \hat{\beta}x_0$ has a large-sample SE given by the estimated square root of

$$\text{var}(\hat{\alpha} + \hat{\beta}x_0) = \text{var}(\hat{\alpha}) + x_0^2 \text{ var}(\hat{\beta}) + 2x_0 \text{ cov}(\hat{\alpha}, \hat{\beta}).$$

A 95% confidence interval for $\text{logit}[\pi(x_0)]$ is $(\hat{\alpha} + \hat{\beta}x_0) \pm 1.96(SE)$. Substituting each endpoint into the inverse transformation $\pi(x_0) = \exp(\text{logit})/[1 + \exp(\text{logit})]$ gives a corresponding interval for $\pi(x_0)$.

5.2.2 Example: Inference for Horseshoe Crab Mating Data

We illustrate logistic regression inferences with the model for the probability that a horseshoe crab has a satellite, with crab width as the predictor. [Table 5.1](#) showed the fit and standard errors. The statistic $z = \hat{\beta}/SE = 0.497/0.102 = 4.89$ provides strong evidence of a positive width effect ($P < 0.0001$). The equivalent Wald chi-squared statistic, $z^2 = 23.89$, has $df = 1$. The maximized log likelihoods equal -112.88 under $H_0: \beta = 0$ and -97.23 for the full model. The likelihood-ratio statistic equals $-2[-112.88 - (-97.23)] = 31.31$, with $df = 1$. This provides even stronger evidence than the Wald test.

The Wald 95% confidence interval for β is $0.497 \pm 1.96(0.102)$, or $(0.298, 0.697)$. [Table 5.1](#) reports a likelihood-ratio confidence interval of $(0.308, 0.709)$, based on the profile likelihood function. The confidence interval for the effect on the odds per 1-cm increase in width equals $(e^{0.308}, e^{0.709}) = (1.36, 2.03)$. We infer that a 1-cm increase in width has at least a 36% increase and at most a doubling in the odds of a satellite.

Most software for logistic regression also can report estimates and confidence intervals for $\pi(x)$ (for examples, see the text website). Consider this for crabs of width $x = 26.5$, which is near the mean width. The estimated logit is $-12.351 + 0.497(26.5) = 0.826$, and $\hat{\pi}(x) = 0.695$. Software reports

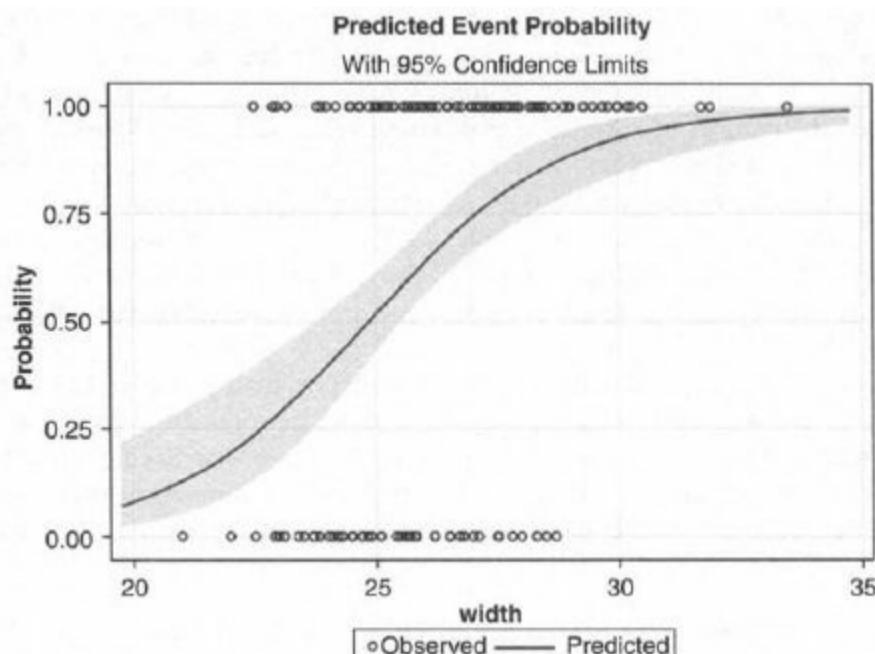
$$\widehat{\text{var}}(\hat{\alpha}) = 6.91023, \quad \widehat{\text{var}}(\hat{\beta}) = 0.01035, \quad \widehat{\text{cov}}(\hat{\alpha}, \hat{\beta}) = -0.26685,$$

from which

$$\widehat{\text{var}}\{\text{logit}[\hat{\pi}(x)]\} = 6.91023 + x^2(0.01035) + 2x(-0.26685).$$

At $x = 26.5$ this is 0.0356, so the 95% confidence interval for $\text{logit}[\pi(26.5)]$ equals $0.826 \pm (1.96)\sqrt{0.0356}$, or $(0.456, 1.196)$. This translates to the interval $(0.61, 0.77)$ for the probability of satellites (e.g., $\exp(0.456)/[1 + \exp(0.456)] = 0.61$). Since $\text{corr}(\hat{\alpha}, \hat{\beta})$ is near 1.0, for better computational precision, fit the model using predictor $x^* = x - 26.5$, so that $\hat{\alpha}$ and its SE are the estimated logit and its SE . [Figure 5.4](#) plots the confidence bands around the prediction equation for $\pi(x)$ as a function of x . Hauck (1983) gave alternative bands for which the confidence coefficient applies simultaneously to all possible predictor values.

[Figure 5.4](#) Prediction equation and 95% confidence bands (from SAS PROC LOGISTIC) for probability of satellite as a function of width.



We could ignore the model fit and simply use sample proportions (i.e., the saturated model) to estimate such probabilities. Six female crabs in the sample had $x = 26.5$, and four of them had satellites. The sample proportion estimate at $x = 26.5$ is $\hat{\pi} = 4/6 = 0.67$, similar to the model-based estimate. The 95% score confidence interval (Section 1.4.2) based on these six observations alone equals $(0.30, 0.90)$.

When the logistic regression model truly holds, the model-based estimator of a probability is considerably better than the sample proportion. The model has only two parameters to estimate, whereas the saturated model has a separate parameter for every distinct value of x . For instance, at $x = 26.5$, software reports $SE = 0.04$ for the model-based estimate 0.695, whereas the SE is $\sqrt{\hat{\pi}(1 - \hat{\pi})/n} = \sqrt{(0.67)(0.33)/6} = 0.19$ for the sample proportion of 0.67 with only 6 observations. The 95% confidence intervals are (0.61, 0.77) using the model versus (0.30, 0.90) using the sample proportion. Instead of using only 6 observations, the model uses the information that all 173 observations provide in estimating the two model parameters. The result is a much more precise estimate.

Reality is a bit more complicated. In practice, the model is not *exactly* the true relationship between $\pi(x)$ and x . However, if it approximates the true probabilities decently, its estimator still tends to be closer than the sample proportion to the true value. The model smooths the sample data, somewhat dampening the observed variability. The resulting estimators tend to be better unless each sample proportion is based on an extremely large sample. Section 5.3.10 discusses this advantage of using models.

5.2.3 Checking Goodness of Fit: Grouped and Ungrouped Data

In practice, there is no guarantee that a certain logistic regression model fits the data well. For any type of binary data, one way to detect lack of fit uses a likelihood-ratio test to compare the model to more complex ones. A more complex model might contain a nonlinear effect. Models with multiple predictors would consider interaction. If more complex models do not fit better, this provides some assurance that the model chosen is reasonable.

Other approaches to detecting lack of fit search for *any* way that the model fails. This is simplest when the explanatory variables are solely categorical, as we'll illustrate in Section 5.4.3. At each setting of x , multiplying the estimated probabilities of the two outcomes by the number of subjects at that setting yields estimated expected frequencies for $y = 0$ and $y = 1$. These are *fitted values*. The test of the model compares the observed counts and fitted values using a Pearson X^2 or likelihood-ratio G^2 statistic. For a fixed number of settings, as the fitted counts increase, X^2 and G^2 have limiting chi-squared null distributions. The degrees of freedom, called the *residual df* for the model, subtract the number of parameters in the model from the number of parameters in the saturated model (i.e., the number of settings of x).

The reason for the restriction to categorical predictors for a global test of fit relates to the distinction that we mentioned in Section 4.5.3 between grouped and ungrouped data for binomial models. The saturated model differs in the two cases. An asymptotic chi-squared distribution for the deviance results as $n \rightarrow \infty$ with a fixed number of parameters in that model and hence a fixed number of settings of predictor values (i.e., *grouped* data).

5.2.4 Example: Model Goodness of Fit for Horseshoe Crab Data

We illustrate with a goodness-of-fit analysis for the model using $x = \text{width}$ to predict the probability that a female crab has a satellite. One way to check it compares it to a more complex model, such as the model containing a quadratic term or linear spline. With width centered at 0 by subtracting its mean of 26.3, the quadratic model has fit

$$\text{logit}[\hat{\pi}(x)] = 0.618 + 0.533(x - \bar{x}) + 0.040(x - \bar{x})^2.$$

The quadratic estimate has $SE = 0.046$. There is not much evidence to support adding that term. The likelihood-ratio statistic for testing that the true coefficient of x^2 is 0 equals 0.83 ($df = 1$).

We next evaluate overall goodness of fit. Width takes 66 distinct values for the 173 crabs, with few observations at most widths. We can view the data as a 66×2 contingency table. The two cells in each row count the number of crabs with satellites and the number of crabs without satellites, at that width. The chi-squared theory for X^2 and G^2 applies when the number of levels of x is fixed, and the number of observations at each level grows. Although we grouped the data using the distinct width values rather than using 173 separate binary responses, this theory is violated here in two ways. First, most fitted counts are very small. Second, when more data are collected, additional width values would occur, so the contingency table would contain more cells rather than a fixed number. Because of this, X^2 and G^2 for logistic regression models with continuous or nearly continuous predictors do not have approximate chi-squared distributions. Normal approximations can be more appropriate (see Section 10.6.4 for references), but no such method has become as popular as methods presented next.

5.2.5 Checking Goodness of Fit with Ungrouped Data by Grouping

As just noted, with ungrouped data or with continuous or nearly continuous predictors, X^2 and G^2 do not have limiting chi-squared distributions. They are still useful for comparing models, as done above for checking a quadratic term. Also, we can apply them in an approximate manner to grouped observed and fitted values for a partition of the space of x values.

[Table 5.2](#) uses the groupings of [Table 4.4](#), giving an 8×2 table. In each width category, the fitted value for a “yes” response is the sum of the estimated probabilities $\hat{\pi}(x)$ for all crabs having width in that category; the fitted value for a “no” response is the sum of $1 - \hat{\pi}(x)$ for those crabs. The fitted values are then much larger. Then, X^2 and G^2 have better validity, although the chi-squared theory still is not perfect because $\pi(x)$ is not constant in each category. Their values are $X^2 = 5.3$ and $G^2 = 6.2$. [Table 5.2](#) has eight binomial samples, one for each width setting; the model has two parameters, so $df = 8 - 2 = 6$. Neither X^2 nor G^2 shows evidence of lack of fit ($P > 0.4$). Thus, we can feel more comfortable about using the model for the original ungrouped data.

Table 5.2 Grouping of Observed and Fitted Values for Fit of Logistic Regression Model to Horseshoe Crab Data

Width (cm)	Number Yes	Number No	Fitted Yes	Fitted No
<23.25	5	9	3.64	10.36
23.25–24.25	4	10	5.31	8.69
24.25–25.25	17	11	13.78	14.22
25.25–26.25	21	18	24.23	14.77
26.25–27.25	15	7	15.94	6.06
27.25–28.25	20	4	19.38	4.62
28.25–29.25	15	3	15.65	2.35
>29.25	14	0	13.08	0.92

As the number of explanatory variables increases, this strategy loses effectiveness. Simultaneous grouping of values for each variable can produce a contingency table with a large number of cells, most of which have very small counts.

Regardless of the number of explanatory variables, we can partition observed and fitted values according to the estimated probabilities of success using the original ungrouped data. One common approach forms the groups in the partition so they have approximately equal size. With 10 groups, the first pair of observed counts and corresponding fitted counts refers to the $n/10$ observations having the highest estimated probabilities, the next pair refers to the $n/10$ observations having the second decile of estimated probabilities, and so on. Each group has an observed count of subjects with each outcome and a fitted value for each outcome. The fitted value for an outcome is the sum of the estimated probabilities for that outcome for all observations in that group.

This construction is the basis of a test due to Hosmer and Lemeshow (1980). They proposed a Pearson statistic comparing the observed and fitted counts for this partition. Let y_{ij} denote the binary outcome for observation j in group i of the partition, $i = 1, \dots, g$, $j = 1, \dots, n_i$. Let $\hat{\pi}_{ij}$ denote the corresponding fitted probability for the model fitted to the ungrouped data. Their statistic equals

$$\sum_{i=1}^g \frac{(\sum_j y_{ij} - \sum_j \hat{\pi}_{ij})^2}{(\sum_j \hat{\pi}_{ij})[1 - (\sum_j \hat{\pi}_{ij})/n_i]}.$$

When many observations have the same estimated probability, there is some arbitrariness in forming the groups, and different software may report somewhat different values. This statistic does not have a limiting chi-squared distribution, because the observations in a group do not share a common success probability and thus are not identical trials. However, Hosmer and Lemeshow noted that when the number of distinct patterns of covariate values equals the sample size, the null distribution is approximated by chi-squared with $df = g - 2$.

For the logistic regression fitted to the horseshoe crab data with continuous width predictor, the

Hosmer–Lemeshow statistic with $g = 10$ groups equals 3.5, with $df = 8$. It also indicates a decent fit.

In summary, the X^2 and G^2 goodness-of-fit tests work well when n is large relative to the number of distinct covariate patterns, whereas the Hosmer–Lemeshow test works well when the number of distinct covariate patterns is large. Unfortunately, like other proposed global fit statistics, the Hosmer–Lemeshow statistic does not have good power for detecting particular types of lack of fit (Hosmer et al. 1997). One example is when the correct model has an interaction between a binary and continuous covariate but the chosen model has only the continuous covariate. Tsiatis (1980) suggested an alternative goodness-of-fit test that partitions values for the explanatory variables into a set of regions and adds an indicator variable to the model for each region. The test statistic compares the fit of this model to the simpler one, testing that the extra parameters are not needed. Alternatively, one could use a bootstrap method to evaluate fit. Azzalini et al. (1989) used the parametric bootstrap to evaluate the distance between the logistic model fit and a nonparametric smoothing of the data (to be introduced in Section 7.4.2); the bootstrap simulations estimated the proportion of times that a likelihood-ratio form of statistic is larger than observed. In any case, a large value of any global fit statistic merely indicates *some* lack of fit but provides no insight about its nature. The approach of comparing the working model to a more complex one is more useful from a scientific perspective, since it searches for lack of fit of a particular type.

For any approach to checking fit, when the fit is poor, diagnostic measures describe the influence of individual observations on the model fit and highlight reasons for the inadequacy. We discuss these in Section 6.2.1.

5.2.6 Wald Inference Can Be Suboptimal

Wald, likelihood-ratio, and score methods of inference usually give similar results for large samples. Each method of inference can also produce small-sample confidence intervals and tests. We defer discussion of this until Sections 7.3, 16.5, and 16.6.

Although these methods usually give similar results, the Wald method has two disadvantages compared with the likelihood-ratio and score methods. First, its results depend on the scale for the parameterization. To illustrate, suppose that Y has a $\text{bin}(n, \pi)$ distribution. For the model, $\text{logit}(\pi) = \alpha$, consider testing $H_0: \alpha = 0$ (i.e., $\pi = 0.50$). From Section 3.1.6, the asymptotic variance of $\hat{\alpha} = \text{logit}(\hat{\pi})$ (with $\hat{\pi} = y/n$) is $[n\pi(1 - \pi)]^{-1}$. The Wald chi-squared test statistic is $[\text{logit}(\hat{\pi})]^2[n\hat{\pi}(1 - \hat{\pi})]$. On the proportion scale, the Wald statistic is $(\hat{\pi} - 0.50)^2[n/\hat{\pi}(1 - \hat{\pi})]$. These are not the same. For example, when $\hat{\pi}$ is near 0 or 1 (so $|\hat{\alpha}|$ is large), the ratio of the Wald statistic on the logit scale to the Wald statistic on the proportion scale approaches 0 as n increases. Evaluations reveal that the logit-scale statistic tends to be too conservative and the proportion-scale statistic tends to be too liberal.

This behavior of the Wald statistic for the logit reflects another disadvantage. When a true effect is relatively large, the Wald test is not as powerful as the likelihood-ratio and score test and can even show aberrant behavior (Hauck and Donner 1977). For the single-binomial case just described, for example, suppose $n = 25$. We would regard $y = 24$ as stronger evidence against H_0 than $y = 23$, yet the logit Wald statistic equals 9.7 when $y = 24$ and 11.0 when $y = 23$. For comparison, the likelihood-ratio statistics are 26.3 and 20.7.

More generally, Hauck and Donner showed that for fixed sample size, the Wald statistic for testing $H_0: \beta = 0$ in the logistic model eventually starts decreasing and actually converges toward 0 as $\hat{\beta}$ grows unboundedly. A similar result holds for logistic models with multiple predictors.

5.3 LOGISTIC MODELS WITH CATEGORICAL PREDICTORS

Like ordinary regression, logistic regression extends to include qualitative explanatory variables, often called *factors*, as first noted by Dyke and Patterson (1952). We use indicator variables to do this.

5.3.1 ANOVA-Type Representation of Factors

For simplicity, we first consider a single factor X , with I categories. In row i of the $I \times 2$ table, let y_i be the number of outcomes in the first column (successes) out of n_i trials. We treat y_i as binomial with parameter π_i .

The logistic regression model with a single factor as a predictor is

$$(5.4) \quad \log \frac{\pi_i}{1 - \pi_i} = \alpha + \beta_i.$$

The higher β_i is, the higher the value of π_i . The right-hand side of (5.4) resembles the model formula for means in one-way ANOVA.

As in ANOVA, the factor has as many parameters $\{\beta_i\}$ as categories. Unless we delete α from the model, one β_i is redundant. One β_i can be set to 0, say, $\beta_I = 0$ for the last category. If the values do not satisfy this, we can recode so that it is true. For instance, set $\tilde{\beta}_i = \beta_i - \beta_I$ and $\tilde{\alpha} = \alpha + \beta_I$, which satisfy $\tilde{\beta}_I = 0$. Then

$$\text{logit}(\pi_i) = \alpha + \beta_i = (\tilde{\alpha} - \tilde{\beta}_I) + (\tilde{\beta}_i + \beta_I) = \tilde{\alpha} + \tilde{\beta}_i,$$

where the newly defined parameters satisfy the constraint. When $\beta_I = 0$, α equals the logit in row I , and β_i is the difference between the logits in rows i and I . Thus, β_i equals the log odds ratio for that pair of rows.

For any $\{\pi_i > 0\}$, $\{\beta_i\}$ exist such that model (5.4) holds. The model has as many parameters (I) as binomial observations and is *saturated*. When a factor has *no* effect, $\beta_1 = \beta_2 = \dots = \beta_I$. Since this is equivalent to $\pi_1 = \dots = \pi_I$, this case corresponds to statistical independence of X and Y .

5.3.2 Indicator Variables Represent a Factor

An equivalent expression of model (5.4) uses indicator variables. Let $x_i = 1$ for observations in row i and $x_i = 0$ otherwise, $i = 1, \dots, I - 1$. The model is

$$\text{logit}(\pi_i) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{I-1} x_{I-1}.$$

This accounts for parameter redundancy by not forming an indicator variable for category I . The constraint $\beta_I = 0$ corresponds to this choice of indicator variables. The category to exclude for an indicator variable is arbitrary. Some software sets $\beta_1 = 0$; this corresponds to a model with indicator variables for categories 2 through I , but not category 1.

Another way to impose constraints sets $\sum_i \beta_i = 0$. When X has $I = 2$ categories, then $\beta_1 = -\beta_2$. This results from *effect coding* for an indicator variable, $x = 1$ in category 1 and $x = -1$ in category 2.

The same substantive results about estimable effects occur for any coding scheme. For model (5.4), regardless of the constraint for $\{\beta_i\}$, the linear predictor values $\{\hat{\alpha} + \hat{\beta}_i\}$ and hence $\{\hat{\pi}_i\}$ are the same. The differences $\hat{\beta}_a - \hat{\beta}_b$ for pairs (a, b) of categories of X are identical and represent estimated log odds ratios. Thus, $\exp(\hat{\beta}_a - \hat{\beta}_b)$ is the estimated odds of success in category a of X divided by the estimated odds of success in category b of X . Reparameterizing a model may change parameter estimates but does not change the model fit or the effects of interest.

The value β_i or $\hat{\beta}_i$ for a single category is irrelevant. Different constraint systems result in different values. For a binary predictor, for instance, using indicator variables with reference value $\beta_2 = 0$, the log odds ratio equals $\beta_1 - \beta_2 = \beta_1$; by contrast, for effect coding with ± 1 indicator variable and hence $\beta_1 + \beta_2 = 0$, the log odds ratio equals $\beta_1 - \beta_2 = \beta_1 - (-\beta_1) = 2\beta_1$. A parameter or its estimate makes sense only by comparison with one for another category.

5.3.3 Example: Alcohol and Infant Malformation Revisited

We return now to [Table 3.8](#) from the study of maternal alcohol consumption and child's congenital malformations, shown again in [Table 5.3](#). For model [\(5.4\)](#), we treat malformation status as the response and alcohol consumption as an explanatory factor. Regardless of the constraint for $\{\beta_i\}$, the model is saturated and $\{\hat{\alpha} + \hat{\beta}_i\}$ are the sample logits, reported in [Table 5.3](#). For instance,

Table 5.3 Sample Logits and Proportion of Malformation for [Table 3.8](#), with Fitted Proportions for Linear Logit Model

Alcohol Consumption	Malformation		Sample Logit	Proportion Malformed	
	Present	Absent		Observed	Fitted
0	48	17,066	-5.87	0.0028	0.0026
<1	38	14,464	-5.94	0.0026	0.0030
1–2	5	788	-5.06	0.0063	0.0041
3–5	1	126	-4.84	0.0079	0.0091
≥6	1	37	-3.61	0.0263	0.0231

$$\text{logit}(\hat{\pi}_1) = \hat{\alpha} + \hat{\beta}_1 = \log(48/17,066) = -5.87.$$

For the coding that constrains $\beta_5 = 0$, $\hat{\alpha} = -3.61$ and $\hat{\beta}_1 = -2.26$. For the coding $\beta_1 = 0$, $\hat{\alpha} = -5.87$.

[Table 5.3](#) shows that except for the slight reversal between the first and second categories of alcohol consumption, the sample logits and hence the sample proportions of malformation cases increase as alcohol consumption increases.

The simpler model with all $\beta_i = 0$ specifies independence. For it, $\hat{\alpha}$ equals the logit for the overall sample proportion of malformations, which is $\log(93/32,481) = -5.86$. To test H_0 : independence (df = 4), the Pearson statistic [\(3.13\)](#) is $X^2 = 12.1$ ($P = 0.02$), and the likelihood-ratio statistic [\(3.11\)](#) is $G^2 = 6.2$ ($P = 0.19$). These provide mixed signals. [Table 5.3](#) has a mixture of very small, moderate, and extremely large counts. Even though $n = 32,574$, the null sampling distributions of X^2 or G^2 may not be close to chi-squared. The P -values using the exact conditional distributions of X^2 and G^2 (Section 16.5.2) are 0.03 and 0.13. These are closer, but still give differing evidence. In any case, these statistics ignore the ordinality of alcohol consumption. The sample suggests that malformations may tend to be more likely with higher alcohol consumption. The first two proportions are similar and the next two are also similar, however, and either of the last two proportions changes substantially with the addition or deletion of one malformation case.

5.3.4 Linear Logit Model for $I \times 2$ Contingency Tables

Model (5.4) is invariant to the ordering of categories, so it treats the explanatory factor as nominal. For ordered factor categories, other models are more parsimonious, yet more complex than the independence model. For instance, let (x_1, x_2, \dots, x_I) be scores that describe distances between categories of X . When we expect a monotone effect of X on Y , it is natural to fit the *linear logit model*

$$(5.5) \text{logit}(\pi_i) = \alpha + \beta x_i.$$

The independence model is the special case $\beta = 0$.

The near-monotone increase in the sample logits in Table 5.3 indicates that the linear logit model may fit better than the independence model. As measured, alcohol consumption groups a naturally continuous variable. With scores $(x_1 = 0, x_2 = 0.5, x_3 = 1.5, x_4 = 4.0, x_5 = 7.0)$, the last score being somewhat arbitrary. Table 5.4 shows results. The estimated multiplicative effect of a unit increase in daily alcohol consumption on the odds of malformation is $\exp(0.317) = 1.37$. Table 5.3 shows the observed and fitted proportions of malformation. The model seems to fit well, as statistics comparing observed and fitted counts are $G^2 = 1.95$ and $X^2 = 2.05$, with $df = 3$.

Table 5.4 Software Output (Based on SAS) for Linear Logit Model Fitted to Table 5.3 on Infant Malformation and Alcohol Consumption

Criteria For Assessing Goodness Of Fit					
Criterion	DF	Value			
Deviance	3	1.9487			
Pearson Chi-Square	3	2.0523			
Log Likelihood		-635.5968			

Parameter	Estimate	Std Error	Likelihood-Ratio	Wald	
				Chi-Sq	Pr>ChiSq
Intercept	-5.9605	0.1154	-6.1930 -5.7397	2666.41	<.0001
alcohol	0.3166	0.1254	0.0187 0.5236	6.37	0.0116

5.3.5 Cochran–Armitage Trend Test

Armitage (1955) and Cochran (1954) were among the first to emphasize the importance of utilizing ordered categories in a contingency table. For $I \times 2$ tables with ordered rows and I independent bin(n_i, π_i) variates $\{y_i\}$, they proposed a trend statistic for testing independence by partitioning the Pearson statistic for that hypothesis. They used a linear probability model,

$$(5.6) \quad \pi_i = \alpha + \beta x_i,$$

fitted by ordinary least squares. The null hypothesis of independence is $H_0: \beta = 0$. Let $\bar{x} = \sum_i n_i x_i / n$. Let $p_i = y_i / n_i$, and let $p = (\sum_i y_i) / n$ denote the overall proportion of successes. The prediction equation is

$$\hat{\pi}_i = p + b(x_i - \bar{x}),$$

where

$$b = \frac{\sum_i n_i(p_i - p)(x_i - \bar{x})}{\sum_i n_i(x_i - \bar{x})^2}.$$

Denote the Pearson statistic for testing independence by $X^2(I)$. We express $X^2(I)$ in terms of variation among the I sample proportions by

$$X^2(I) = \frac{1}{p(1-p)} \sum_i n_i(p_i - p)^2.$$

Reported by Fisher (1934) and attributed to A. E. Brandt and G. W. Snedecor, this is referred to as the *Brandt–Snedecor formula*. It generalizes the equality in 2×2 tables between X^2 and the square of the pooled two-sample z -statistic (3.12). Cochran (1954) noted that this Pearson formula decomposes into

$$X^2(I) = z^2 + X^2(L),$$

where

$$(5.7) \quad \begin{aligned} X^2(L) &= \frac{1}{p(1-p)} \sum_i n_i(p_i - \hat{\pi}_i)^2, \\ z^2 &= \frac{b^2}{p(1-p)} \sum_i n_i(x_i - \bar{x})^2 = \left[\frac{\sum_i (x_i - \bar{x}) y_i}{\sqrt{p(1-p) \sum_i n_i (x_i - \bar{x})^2}} \right]^2. \end{aligned}$$

When the linear probability model holds, $X^2(L)$ is asymptotically chi-squared with $df = I - 2$. It tests the fit of the model. The statistic z^2 , with $df = 1$, tests $H_0: \beta = 0$ for the linear trend (5.6) in the proportions. The test of independence using this statistic is called the *Cochran–Armitage trend test*.

This statistic relates to the correlation-based statistic M^2 introduced in (3.16) in Section 3.4.1 to test for a linear trend in an $I \times J$ table; namely, $z^2 = nr^2 = [n/(n-1)]M^2$. See Yates (1948) and Mantel (1963). When $I = 2$, then $X^2(L) = 0$ and $z^2 = X^2(I)$.

The Cochran–Armitage trend test seems unrelated to the linear logit model. However, this test statistic is equivalent to the score statistic for testing $H_0: \beta = 0$ in that model. In fact, Tarone and Gart (1980) showed that the score test for a binary linear trend model does not depend on the link function. Thus, this trend test is locally asymptotically efficient for both linear and logistic alternatives for $P(Y = 1)$. See Cox (1958a) for related remarks. Gross (1981) showed that when the linear logit model holds but we use an incorrect set of scores, the local asymptotic relative efficiency for testing independence using the statistic with those scores equals the square of the Pearson correlation between the true and the incorrect scores.

5.3.6 Example: Alcohol and Infant Malformation Revisited

For [Table 5.3](#) on alcohol consumption and infant malformation, $X^2(I) = 12.08$. Using the scores (0, 0.5, 1.5, 4.0, 7.0) as in the linear logit model, the Cochran–Armitage trend test has $z^2 = 6.57$ (P -value = 0.010). The test suggests strong evidence of a positive slope. In addition,

$$X^2(I) = 12.08 = 6.57 + 5.51,$$

where $X^2(L) = 5.51$ ($df = 3$) shows only slight evidence of departure of the proportions from linearity. The trend test result is nearly identical to the test using $M^2 = (n - 1)r^2$ based on the sample correlation of $r = 0.0142$ for $n = 32, 573$. For the chosen scores, the correlation seems weak. However, r has limited use as a descriptive measure for tables that are highly discrete and unbalanced.

The Cochran–Armitage trend test (i.e., the score test) usually gives results similar to the Wald or likelihood-ratio test of $H_0: \beta = 0$ in the linear logit model. The asymptotics work well even for quite small n when $\{n_i\}$ are equal and $\{x_i\}$ are equally spaced. With [Table 5.3](#), the Wald statistic equals $(\hat{\beta}/SE)^2 = (0.3166/0.1254)^2 = 6.37$ ($P = 0.012$) and the likelihood-ratio statistic equals 4.25 ($P = 0.039$). Here, however, the highly unbalanced counts suggest that it is best not to use the Wald approach for testing or for interval estimation. The profile likelihood 95% confidence interval of (0.02, 0.52) for β reported in [Table 5.4](#) is preferable to the Wald interval of $0.317 \pm 1.96(0.125) = (0.07, 0.56)$. The sample size in the last row is relatively small, and the single “present” observation in that row is highly influential. P -values depend dramatically on whether that observation is included in the analysis (Exercise 5.10).

5.3.7 Using Directed Models Can Improve Inferential Power

When contingency tables have ordered categories, in Section 3.4 we showed that tests that utilize the ordering can have improved power. Testing independence against a linear trend alternative in a linear logit model is a way to do this. In this section we present the reason for these power improvements.

In an $I \times 2$ contingency table for I binomial variates with parameters $\{\pi_i\}$, H_0 : independence states $\text{logit}(\pi_i) = \alpha$. The ordinary G^2 and X^2 statistics of Section 3.2.1 refer to the general alternative,

$$\text{logit}(\pi_i) = \alpha + \beta_i,$$

which is saturated. They test $H_0: \beta_1 = \beta_2 = \dots = \beta_I$ in that model, with $\text{df} = (I - 1)$. Their general alternative treats both classifications as nominal. Denote these test statistics as $G^2(I)$ and $X^2(I)$. Note that $G^2(I)$ is the likelihood-ratio statistic $G^2(M_0|M_1) = -2(L_0 - L_1)$ for comparing the saturated model M_1 with the independence (I) model M_0 .

Ordinal test statistics refer to narrower, often more relevant, alternatives. With ordered rows, an example is a test of $H_0: \beta = 0$ in the linear logit model, $\text{logit}(\pi_i) = \alpha + \beta x_i$. The likelihood-ratio statistic $G^2(I|L) = G^2(I) - G^2(L)$ compares the linear logit model and the independence model. When a test statistic focuses on a single parameter, such as β in that model, it has $\text{df} = 1$. Now, df equals the mean of the chi-squared distribution. A large test statistic with $\text{df} = 1$ falls farther out in its right-hand tail than a comparable value of $X^2(I)$ or $G^2(I)$ with $\text{df} = (I - 1)$. Thus, it has a smaller P -value.

5.3.8 Noncentral Chi-Squared Distribution and Power for Narrower Alternatives

To compare power of $G^2(I|L)$ and $G^2(I)$, it is necessary to compare their nonnull sampling distributions. When H_0 is false, their distributions are approximately *noncentral chi-squared*. This distribution, introduced by R. A. Fisher in 1928, arises from the following construction: If $Z_i \sim N(\mu_i, 1)$, $i = 1, \dots, v$, and if Z_1, \dots, Z_v are independent, $\sum_i Z_i^2$ has the noncentral chi-squared distribution with $df = v$ and *noncentrality parameter* $\lambda = \sum_i \mu_i^2$. Its mean is $v + \lambda$ and its variance is $2(v + 2\lambda)$. The ordinary (central) chi-squared distribution, which occurs when H_0 is true, has $\lambda = 0$.

Let $X_{v,\lambda}^2$ denote a noncentral chi-squared random variable with $df = v$ and noncentrality λ . A fundamental result for chi-squared analyses is that, for fixed λ ,

$$P[X_{v,\lambda}^2 > \chi_v^2(\alpha)] \text{ increases as } v \text{ decreases.}$$

That is, the power for rejecting H_0 at a fixed α -level increases as the df of the test decreases (Das Gupta and Perlman 1974). For fixed v , the power equals α when $\lambda = 0$, and it increases as λ increases. The inverse relation between power and df suggests that focusing the noncentrality on a statistic having a small df value can improve power.

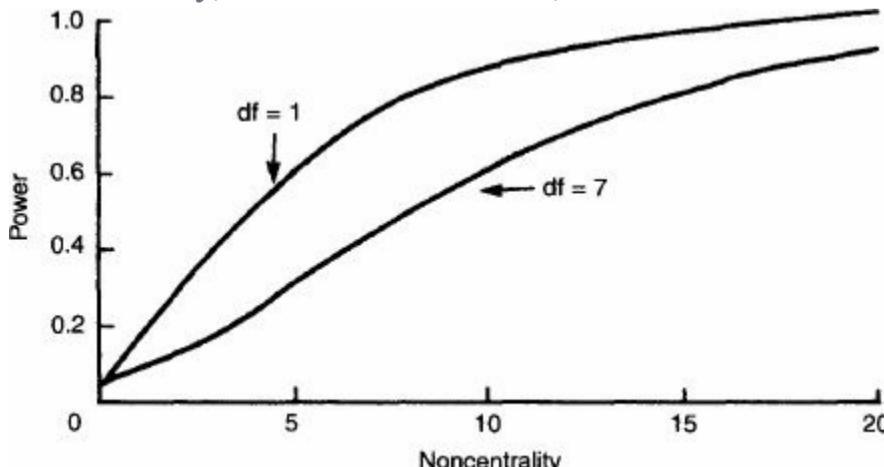
Suppose that an explanatory variable has, at least approximately, a linear effect on $\text{logit}[P(Y=1)]$. To test independence with reasonable power, it is then sensible to use a statistic based on the linear logit model, using the likelihood-ratio statistic $G^2(I|L)$, the Wald statistic $z = \beta/\text{SE}$, and the Cochran–Armitage (score) statistic. When is $G^2(I|L)$ more powerful than $G^2(I)$? The statistics satisfy

$$G^2(I) = G^2(I|L) + G^2(L),$$

where $G^2(L)$ tests goodness of fit of the linear logit model. When the linear logit model holds, $G^2(L)$ has an asymptotic chi-squared distribution with $df = I - 2$; then if $\beta \neq 0$, $G^2(I)$ and $G^2(I|L)$ both have approximate noncentral chi-squared distributions with the same noncentrality. Whereas $df = I - 1$ for $G^2(I)$, $df = 1$ for $G^2(I|L)$. Thus, $G^2(I|L)$ is more powerful, because it uses fewer degrees of freedom.

When the linear logit model does not hold, $G^2(I)$ has greater noncentrality than $G^2(I|L)$, the discrepancy increasing as the model fits more poorly. However, when the model approximates reality fairly well, usually $G^2(I|L)$ is still more powerful. That test's df value of 1 more than compensates for its loss in noncentrality. The closer the true relationship is to the linear logit, the more nearly $G^2(I|L)$ captures the same noncentrality as $G^2(I)$, and the more powerful it is compared with $G^2(I)$. To illustrate, [Figure 5.5](#) plots power as a function of noncentrality when $df = 1$ and 7. When the noncentrality of a test having $df = 1$ is at least about half that of a test having $df = 7$, the test with $df = 1$ is more powerful. The linear logit model then helps detect a key component of an association. As Mantel (1963) argued in a similar context, “that a linear regression is being tested does not mean that an assumption of linearity is being made. Rather it is that test of a linear component of regression provides power for detecting any progressive association which may exist.”

[Figure 5.5](#) Power and noncentrality, for $df = 1$ and $df = 7$, when $\alpha = 0.05$.



The improved power for the linear trend statistic results from sacrificing power in other cases. The

$G^2(I)$ test can have greater power than $G^2(I|L)$ when the linear logit model describes the true relationship very poorly.

5.3.9 Example: Skin Damage and Leprosy

[Table 5.5](#) refers to an experiment on the use of sulfones and streptomycin drugs in the treatment of leprosy. The degree of infiltration at the start of the experiment measures a type of skin damage. The response is the change in the overall clinical condition of the patient after 48 weeks of treatment. We use response scores (-1, 0, 1, 2, 3). The question of interest is whether subjects with high infiltration changed differently from those with low infiltration.

Table 5.5 Change in Clinical Condition by Degree of Infiltration

Clinical Change	Degree of Infiltration		Proportion High
	High	Low	
Worse	1	11	0.08
Stationary	13	53	0.20
Slight improvement	16	42	0.28
Moderate improvement	15	27	0.36
Marked improvement	7	11	0.39

Source: Reprinted with permission from the Biometric Society (Cochran 1954).

The test $G^2(I) = 7.28$ ($df = 4$) does not show much evidence of association ($P = 0.12$), but it ignores that the clinical change response variable is ordinal. It seems natural to compare the mean change for the two infiltration levels. Cochran (1954) and Yates (1948) noted that this analysis is identical to a trend test treating the binary variable as the response. In fact, the sample proportion of high infiltration increases monotonically as the clinical change improves. The test of $H_0: \beta = 0$ in the linear logit model has $G^2(I|L) = 6.65$, with $df = 1$ ($P = 0.01$). It gives strong evidence of more positive clinical change at the higher level of infiltration. Using the ordering by decreasing df from 4 to 1 pays a strong dividend. In addition, $G^2(L) = 0.63$ with $df = 3$ suggests that the linear trend model fits well.

5.3.10 Model Smoothing Improves Precision of Estimation

Using directed alternatives can improve not only *test power*, but also *estimation* of cell probabilities and summary measures. In generic form, let π be true cell probabilities in a contingency table, let p denote sample proportions, and let $\hat{\pi}$ denote model-based ML estimates of π .

When π satisfies a certain model, both $\hat{\pi}$ for that model and p are consistent estimators of π . The model-based estimator $\hat{\pi}$ is better, as its true asymptotic standard error cannot exceed that of p . This happens because of model parsimony: The unsaturated model, on which $\hat{\pi}$ is based, has fewer parameters than the saturated model, on which p is based. In fact, model-based estimators are also more efficient in estimating functions $g(\pi)$ of cell probabilities. For any differentiable function g ,

$$\text{asymp. var}[\sqrt{n}g(\hat{\pi})] \leq \text{asymp. var}[\sqrt{n}g(p)].$$

In Section 16.2.3 we show formulas. The result holds more generally than for categorical data models (Altham 1984), a reason that statisticians prefer parsimonious models.

In reality, of course, a chosen model is unlikely to hold exactly. However, when the model approximates π well, unless n is extremely large, $\hat{\pi}$ is still better than p . Although $\hat{\pi}_i$ is biased, it has smaller variance than p_i , and $\text{MSE}(\hat{\pi}_i) < \text{MSE}(p_i)$ when its variance plus squared bias is smaller than $\text{var}(p_i)$. In Section 3.3.8, for example, we showed that independence-model estimates of cell probabilities in two-way tables can be much better than sample proportions even when that model does not hold.

5.4 MULTIPLE LOGISTIC REGRESSION

Like ordinary regression, logistic regression extends to models with multiple explanatory variables, which can be a mixture of quantitative and qualitative (Cox 1958). The model for $\pi(\mathbf{x}) = P(Y = 1)$ at values $\mathbf{x} = (x_1, \dots, x_p)$ of p predictors is

$$(5.8) \text{ logit}[\pi(\mathbf{x})] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p.$$

The alternative formula, directly specifying $\pi(\mathbf{x})$, is

$$(5.9) \pi(\mathbf{x}) = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}.$$

For qualitative predictors, we use indicator variables for its categories.

The parameter β_j refers to the effect of x_j on the log odds that $Y = 1$, adjusting for the other x_k . For instance, $\exp(\beta_j)$ is the multiplicative effect on the odds of a 1-unit increase in x_j , when we can keep fixed the levels of other x_k .

5.4.1 Logistic Models for Multiway Contingency Tables

When all variables are categorical, a multiway contingency table displays the data. We illustrate ideas with binary predictors X and Z . We treat the sample size at given combinations of X and Z as fixed and regard the two counts on Y at each setting as binomial, with different binomials treated as independent. We let indicator variables x and z take value 1 in the first category and 0 in the second. The model

$$(5.10) \logit[P(Y = 1)] = \alpha + \beta_1 x + \beta_2 z$$

has main effects for X and Z but assumes an absence of interaction. The effect of one factor is the same at each level of the other.

At a fixed level of Z , the effect on the logit of changing categories of X is

$$(5.11) [\alpha + \beta_1(1) + \beta_2 z] - [\alpha + \beta_1(0) + \beta_2 z] = \beta_1.$$

This logit difference equals the difference of log odds, which is the log odds ratio between X and Y , fixing Z . Thus, $\exp(\beta_1)$ is the conditional odds ratio between X and Y . Adjusting for Z , the odds of success when $x = 1$ equal $\exp(\beta_1)$ times the odds when $x = 0$. This conditional odds ratio is the same at each level of z ; that is, there is *homogeneous XY association* (Section 2.3.5). The lack of an interaction term implies a common odds ratio for the partial tables. When $\beta_1 = 0$, that common odds ratio equals 1. Then X and Y are independent in each partial table, or *conditionally independent, given Z* (Section 2.3.4).

Additivity on the logit scale is the usual definition of no interaction for categorical variables. However, it could instead be defined as additivity on some other scale, such as with probit or identity link. Interaction can occur on one scale when there is none on another scale. In some applications, a particular definition may be natural. For instance, theory might assume an underlying normal distribution for Y and predict that the probit is an additive function of predictor effects.

A factor with I categories needs $I - 1$ indicator variables. With I categories for X and K categories for Z , model (5.10) extends to

$$\logit[P(Y = 1)] = \alpha + \beta_1^X x_1 + \cdots + \beta_{I-1}^X x_{I-1} + \beta_1^Z z_1 + \cdots + \beta_{K-1}^Z z_{K-1},$$

where, for example, $z_k = 1$ for observations in category k of Z and $z_k = 0$ otherwise, $k = 1, \dots, K - 1$.

This equation represents effects of X with parameters $\{\beta_i^X\}$ and effects of Z with parameters $\{\beta_k^Z\}$. The X and Z superscripts are merely labels and do not represent powers. This model form applies for any number of categories for X and Z . The parameter β_k^Z , for example, denotes the effect on the logit of classification in category k of Z instead of category K .

An alternative representation of such factors resembles the way that ANOVA factorial models often express them. The equivalent model formula is

$$(5.12) \logit[P(Y = 1)] = \alpha + \beta_i^X + \beta_k^Z.$$

For each factor, one parameter is redundant. Fixing one at 0, such as $\beta_I^X = \beta_K^Z = 0$, represents the category not having its own indicator variable. Conditional independence between X and Y , given Z , corresponds to $\beta_1^X = \beta_2^X = \cdots = \beta_I^X$, whereby $P(Y = 1)$ does not change as i changes, for fixed k .

5.4.2 Example: AIDS and AZT Use

[Table 5.6](#) is from a study on the effects of AZT in slowing the development of AIDS symptoms. In the study, 338 veterans whose immune systems were beginning to falter after infection with HIV were randomly assigned either to receive AZT immediately or to wait until their T cells showed severe immune weakness. [Table 5.6](#) cross-classifies the veterans' race, whether they received AZT immediately, and whether they developed AIDS symptoms during the 3-year study.

Table 5.6 Development of AIDS Symptoms by AZT Use and Race

Race	AZT Use	Symptoms	
		Yes	No
White	Yes	14	93
	No	32	81
Black	Yes	11	52
	No	12	43

Source: *The New York Times*, Feb. 15, 1991.

In model [\(5.10\)](#), we identify X with AZT treatment ($x = 1$ for immediate AZT use, $x = 0$ otherwise) and Z with race ($z = 1$ for whites, $z = 0$ for blacks), for predicting the probability that AIDS symptoms developed. Thus, α is the log odds of developing AIDS symptoms for black subjects without immediate AZT use, β_1 is the increment to the log odds for those with immediate AZT use, and β_2 is the increment to the log odds for white subjects. [Table 5.7](#) shows output. The estimated odds ratio between immediate AZT use and development of AIDS symptoms equals $\exp(-0.7195) = 0.487$. For each race, the estimated odds of symptoms are half as high for those who took AZT immediately. The Wald confidence interval for this effect is $\exp[-0.720 \pm 1.96(0.279)] = (0.28, 0.84)$. Similar results occur for the likelihood-based interval, as shown.

Table 5.7 Software Output (Based on SAS) for Logistic Model with AIDS Symptoms Data

Goodness-of-Fit Statistics				
Criterion	DF	Value	Pr > ChiSq	
Deviance	1	1.3835	0.2395	
Pearson	1	1.3910	0.2382	
Analysis of Maximum Likelihood Estimates				
Parameter	Estimate	Std Error	Wald Chi-Sq	Pr > ChiSq
Intercept	-1.0736	0.2629	16.6705	<.0001
azt	-0.7195	0.2790	6.6507	0.0099
race	0.0555	0.2886	0.0370	0.8476
Obs	race	azt	y	n
1	1	1	14	107
2	1	0	32	113
3	0	1	11	63
4	0	0	12	55
			pihat	lower upper
				0.09897 0.21987
				0.19668 0.34774
				0.08704 0.22519
				0.16953 0.36396
Profile Like. CI for Odds Ratios				
Effect	Estimate	95% Conf Limits		
azt	0.487	0.279 0.835		
race	1.057	0.605 1.884		

The hypothesis of conditional independence of AZT treatment and development of AIDS symptoms, controlling for race, is $H_0: \beta_1 = 0$ in [\(5.10\)](#). The likelihood-ratio statistic comparing the model with the simpler model having $\beta_1 = 0$ equals 6.87 (df = 1), showing evidence of association ($P = 0.01$). The Wald statistic $(\hat{\beta}_1/SE)^2 = (-0.7195/0.279)^2 = 6.65$, shown in the output, provides similar results.

[Table 5.8](#) shows parameter estimates for three ways of defining factor parameters in [\(5.12\)](#): (1) setting the last parameter equal to 0, (2) setting the first parameter equal to 0, and (3) having parameters sum to zero. This corresponds to setting up indicator variables for each category except the last in scheme (1), for each category except the first in scheme (2). In scheme (3), there is also a reference category, and for other categories the indicator is 1 for an observation in the category, -1 for an observation in the reference category, and 0 otherwise. For each coding scheme, at a given

combination of AZT use and race, the estimated probability of developing AIDS symptoms is the same. For instance, the intercept estimate plus the estimate for immediate AZT use plus the estimate for being white is -1.738 for each scheme, so the estimated probability that white veterans with immediate AZT use develop AIDS symptoms equals $\exp(-1.738)/[1 + \exp(-1.738)] = 0.15$. The bottom of [Table 5.7](#) shows point and interval estimates of the probabilities. [Figure 5.6](#) shows a graphical representation of the sample proportions (the four dots) and the point estimates plus and minus a standard error.

Figure 5.6 Estimated effects of AZT use and race on probability of developing AIDS symptoms (dots are sample proportions).

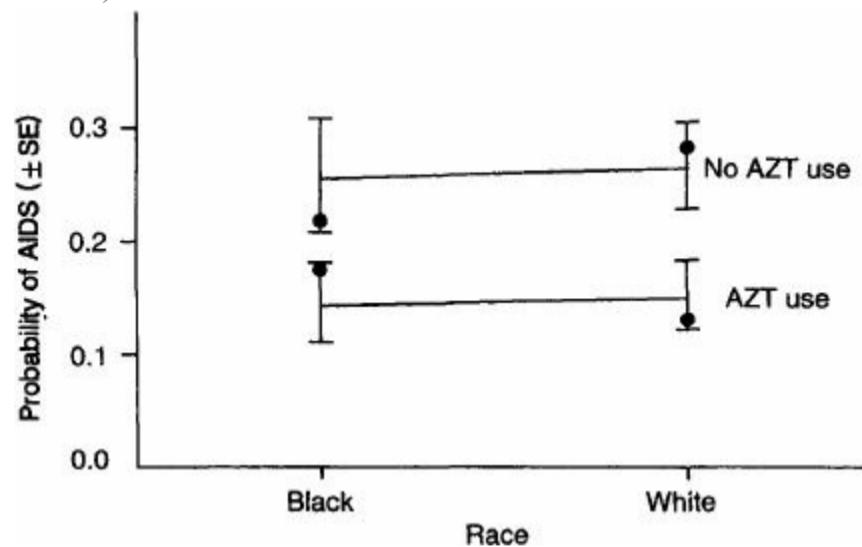


Table 5.8 Parameter Estimates for Logistic Model Fitted to [Table 5.6](#) on AIDS and AZT Use

Parameter	Definition of Parameters		
	Last = Zero	First = Zero	Sum = Zero
Intercept	-1.074	-1.738	-1.406
AZT Yes No	-0.720	0.000	-0.360
	0.000	0.720	0.360
Race White Black	0.055	0.000	0.028
	0.000	-0.055	-0.028

For each coding scheme, $\beta_1^X - \beta_2^X$ is identical and represents the conditional log odds ratio of X with the response, given Z . Here, $\exp(\hat{\beta}_1^X - \hat{\beta}_2^X) = \exp(-0.720) = 0.49$ estimates the common odds ratio between immediate AZT use and AIDS symptoms, for each race.

5.4.3 Goodness of Fit as a Likelihood-Ratio Test

The likelihood-ratio statistic $G^2(M_0|M_1) = -2(L_0 - L_1)$ tests whether certain model parameters are zero, given that M_1 holds, by comparing the log likelihood L_1 for the fitted model M_1 with L_0 for a simpler model M_0 . The goodness-of-fit statistic $G^2(M)$ is a special case in which $M_0 = M$ and M_1 is the saturated model. In testing whether M fits, we test whether *all* parameters in the saturated model but not in M equal zero. The asymptotic df is the difference in the number of parameters in the two models, which is the number of binomials modeled minus the number of parameters in M .

We illustrate by checking the fit of model (5.10) for the AIDS data. For its fit, white veterans with immediate AZT use had estimated probability 0.150 of developing AIDS symptoms during the study. Since 107 white veterans took AZT, the fitted value is $107(0.150) = 16.0$ for developing symptoms and $107(0.850) = 91.0$ for not developing them. Similarly, we can obtain fitted values for all eight cells in Table 5.6. The goodness-of-fit statistics comparing these with the cell counts are $G^2 = 1.38$ and $X^2 = 1.39$. The model has four binomials, one at each combination of AZT use and race. Since it has three parameters, residual df = $4 - 3 = 1$. The small G^2 and X^2 values suggest that the model fits decently ($P > 0.2$).

For model (5.10), the odds ratio between X and Y is the same at each level of Z . The goodness-of-fit test checks this structure. That is, the test also provides a test of homogeneous odds ratios. For Table 5.6, homogeneity is plausible. Since residual df = 1, the more complex model that adds an interaction term and permits the two odds ratios to differ is saturated.

5.4.4 Model Comparison by Comparing Deviances

Let L_S denote the maximized log likelihood for the saturated model. As discussed in Section 4.5.4, the likelihood-ratio statistic for comparing models M_1 and M_0 is

$$G^2(M_0|M_1) = -2(L_0 - L_1) = -2(L_0 - L_S) - [-2(L_1 - L_S)] = G^2(M_0) - G^2(M_1).$$

The test statistic comparing two models is identical to the difference in G^2 goodness-of-fit statistics (deviances) for the two models. To illustrate, consider $H_0: \beta_2 = 0$ for the race effect with the AIDS data. The likelihood-ratio statistic equals 0.04, suggesting that the simpler model is adequate. But this equals $G^2(M_0) - G^2(M_1) = 1.42 - 1.38$, where M_0 is the simpler model with $\beta_2 = 0$.

The model comparison statistic often has an approximate chi-squared null distribution even when separate $G^2(M_i)$ do not. For instance, when at least one predictor is continuous or a contingency table has very small fitted values, the sampling distribution of $G^2(M_i)$ may be far from chi-squared. Nonetheless, if df for the comparison statistic is modest (as in comparing two models that differ by at most a few parameters), the null distribution of $G^2(M_0|M_1)$ is approximately chi-squared.

5.4.5 Example: Horseshoe Crab Satellites Revisited

For the horseshoe crab data, we next use both the female crab's carapace width and color as predictors of $Y =$ whether the crab has at least one satellite ($1 =$ yes, $0 =$ no). Color has five categories: light, medium light, medium, medium dark, dark. It is a surrogate for age, older crabs tending to be darker. The sample contained no light crabs, so our models use only the other four categories. We first treat color as qualitative. The four categories use three indicator variables. The model for the probability that the crab has at least one satellite is

$$(5.13) \text{logit}[P(Y = 1)] = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 x,$$

where $x =$ width in centimeters, and

$c_1 = 1$ for medium-light color, and 0 otherwise,

$c_2 = 1$ for medium color, and 0 otherwise,

$c_3 = 1$ for medium-dark color, and 0 otherwise.

The crab color is dark (category 4) when $c_1 = c_2 = c_3 = 0$. [Table 5.9](#) shows the ML parameter estimates. For instance, for dark crabs, $\text{logit}[\hat{P}(Y = 1)] = -12.715 + 0.468x$; by contrast, for medium-light crabs, $c_1 = 1$, and $\text{logit}[\hat{P}(Y = 1)] = (-12.715 + 1.330) + 0.468x = -11.385 + 0.468x$. At the average width of 26.3 cm, $\hat{P}(y = 1) = 0.399$ for dark crabs and 0.715 for medium-light crabs. The exponentiated difference between two color parameter estimates is an odds ratio comparing those colors. For instance, the difference for medium-light crabs and dark crabs equals 1.330. At any given width, the estimated odds that a medium-light crab has a satellite are $\exp(1.330) = 3.8$ times the estimated odds for a dark crab. At width $x = 26.3$, the odds equal $0.715/0.285 = 2.51$ for a medium-light crab and $0.399/0.601 = 0.66$ for a dark crab, for which $2.51/0.66 = 3.8$.

[Table 5.9](#) Software Output (Based on SAS) for Model with Width and Color Predictors of Whether Horseshoe Crab Has Satellites

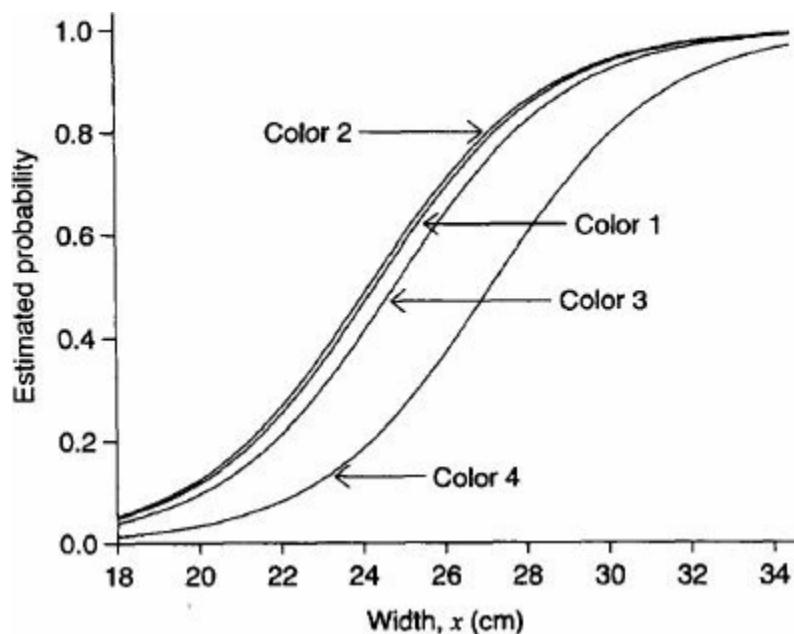
Criteria For Assessing Goodness Of Fit						
Criterion	DF	Value				
Deviance	168	187.4570				
Pearson Chi-Square	168	168.6590				
Log Likelihood		-93.7285				

Parameter	Estimate	Standard Error	Standard	Likelihood-Ratio	95% Confidence Limits	Chi-Square	Pr>ChiSq
			Chi-Square	Pr>ChiSq			
intercept	-12.7151	2.7618	-18.4564	-7.5788	21.20	<.0001	
c1	1.3299	0.8525	-0.2738	3.1354	2.43	0.1188	
c2	1.4023	0.5484	0.3527	2.5260	6.54	0.0106	
c3	1.1061	0.5921	-0.0279	2.3138	3.49	0.0617	
width	0.4680	0.1055	0.2713	0.6870	19.66	<.0001	

To test whether color contributes significantly to model (5.13), we test $H_0: \beta_1 = \beta_2 = \beta_3 = 0$. This states that controlling for width, the probability of a satellite is independent of color. We compare the maximized log-likelihood L_1 for the full model (5.13) to L_0 for the simpler model. The test statistic $-2(L_0 - L_1) = 7.0$ has $df = 3$, the difference between the numbers of parameters in the two models. The chi-squared P -value of 0.07 provides slight evidence of a color effect.

The model assumes a lack of interaction between color and width in their effects. Width has the coefficient of 0.468 for all colors, so the shapes of the curves relating width to $P(Y = 1)$ are identical. [Figure 5.7](#) displays the fitted model. Any one curve equals any other curve shifted to the right or left. The parallelism of curves in the horizontal dimension implies that any two curves never cross. At all width values, color 4 (dark) has a lower estimated probability of a satellite than the other colors. There is a noticeable positive effect of width.

[Figure 5.7](#) Logistic regression model using additive width and color predictors of whether horseshoe crab has satellites.



The more complex model allowing color \times width interaction has three additional terms, the cross-products of width with the color indicator variables. Fitting this model is equivalent to fitting logistic regression with width predictor separately for crabs of each color. Each color then has a different-shaped curve relating width to $P(Y = 1)$, so a comparison of two colors varies according to the width value. The likelihood-ratio statistic comparing the models with and without the interaction terms equals 4.4, with $df = 3$. The evidence of interaction is weak ($P = 0.22$).

5.4.6 Quantitative Treatment of Ordinal Predictor

Color has ordered categories, from lightest to darkest. A simpler model yet treats this predictor as quantitative. Color may have a linear effect, for a set of monotone scores. To illustrate, for scores $c = (1, 2, 3, 4)$ for the color categories, the model

$$(5.14) \text{ logit}[P(Y = 1)] = \alpha + \beta_1 c + \beta_2 x$$

has $\hat{\alpha} = -10.071$, $\hat{\beta}_1 = -0.509$ ($SE = 0.224$) and $\hat{\beta}_2 = 0.458$ ($SE = 0.104$). This shows strong evidence of an effect for each. At a given width, for every one-category increase in color darkness, the estimated odds of a satellite multiply by $\exp(-0.509) = 0.60$.

The likelihood-ratio statistic comparing this fit to the more complex model (5.13) having a separate parameter for each color equals 1.66 ($df = 2$). This statistic tests that the simpler model (5.14) is adequate, given that model (5.13) holds. It tests that when plotted against the color scores, the color parameters in (5.13) follow a linear trend. The simplification seems permissible ($P = 0.44$).

The color parameter estimates in the qualitative-color model (5.13) are $(1.33, 1.40, 1.11, 0)$, the 0 value for the dark category reflecting its lack of an indicator variable. Although these values do not depart significantly from a linear trend, the first three are quite similar compared with the last one. Thus, another potential color scoring for model (5.14) is $(1, 1, 1, 0)$; that is, score = 0 for dark-colored crabs, and score = 1 otherwise. The likelihood-ratio statistic comparing model (5.14) with these binary scores to model (5.13) equals 0.50 ($df = 2$), showing that this simpler model is also adequate. Its fit is

$$(5.15) \text{ logit}[\hat{P}(Y = 1)] = -12.980 + 1.300c + 0.478x,$$

with standard errors 0.526 and 0.104. At a given width, the estimated odds that a lighter-colored crab has a satellite are $\exp(1.300) = 3.7$ times the estimated odds for a dark crab.

In summary, the qualitative-color model, the quantitative-color model with scores $(1, 2, 3, 4)$, and the model with binary color scores $(1, 1, 1, 0)$ all suggest that dark crabs are least likely to have satellites. A much larger sample is needed to determine which color scoring is most appropriate. With moderate-sized samples, it's not unusual for quite different models to be consistent with the data.

5.4.7 Probability-Based and Standardized Interpretations

Although it is natural to interpret logistic regression model parameters as effects on a log odds, some find it difficult to understand odds or odds ratio effects. The simpler interpretation using the instantaneous rate of change in the probability (Section 5.1.1) applies also to multiple predictors. Consider a setting of predictors at which $\hat{P}(Y=1) = \hat{\pi}$. Then, adjusting for the other predictors, as a function of a quantitative predictor x_j , $\hat{\pi}$ has instantaneous rate of change of $\hat{\beta}_j \hat{\pi}(1 - \hat{\pi})$. For instance, at predictor settings at which $\hat{\pi} = 0.50$ for fit (5.15), the approximate effect of a 1-cm increase in width is $(0.478)(0.50)(0.50) = 0.12$. This is considerable, since a 1-cm change in width is less than half a standard deviation.

We could summarize the effect of x_j on the probability scale by averaging the instantaneous rates for the sample. Let x_{ij} denote the value of x_j for subject i and let $\hat{\pi}(x_{i1}, \dots, x_{ip})$ denote the estimate of $P(Y=1)$ at the explanatory variable values for subject i . This summary is

$$\frac{1}{n} \sum_{i=1}^n \hat{\beta}_j \hat{\pi}(x_{i1}, \dots, x_{ip}) [1 - \hat{\pi}(x_{i1}, \dots, x_{ip})].$$

Alternatively, to describe the effect of x_j in a simpler manner not depending on its units, we could set the other predictors at their sample means and compute the estimated probabilities at the smallest and largest x_j values. These are sensitive to outliers, however, so we could instead use the upper and lower quartiles of x_j . For the fit (5.15) with binary color, the sample means are 26.3 for x and 0.873 for c . The lower and upper quartiles of x are 24.9 and 27.7. At $x = 24.9$ and $c = \bar{c}$, $\hat{\pi} = 0.51$. At $x = 27.7$ and $c = \bar{c}$, $\hat{\pi} = 0.80$. The change in $\hat{\pi}$ from 0.51 to 0.80 over the middle 50% of the range of width values reflects a strong width effect. Since c takes only values 0 and 1, we could instead report this effect separately for each. Also, when an explanatory variable is an indicator, it makes sense to report the estimated probabilities at its two values rather than at quartiles, which could be identical. At $\bar{x} = 26.3$, $\hat{\pi} = 0.40$ when $c = 0$ and $\hat{\pi} = 0.71$ when $c = 1$. This color effect, differentiating dark crabs from others, is also substantial.

[Table 5.10](#) summarizes the logistic parameter estimates and some probability comparison effects. It also shows results of the extension of model (5.15), permitting interaction. The estimated width effect is then greater for the lighter-colored crabs. However, the interaction is not significant.

Table 5.10 Summary of Effects in Model (5.15) with Crab Width and Color (Treated as Binary) as Predictors of Presence of Satellites

Variable	Estimate	SE	Comparison	Change in Probability
No interaction model				
Intercept	-12.980	2.727		
Color (0 = dark, 1 = other)	1.300	0.526	(1, 0) at \bar{x}	$0.31 = 0.71 - 0.40$
Width, x (cm)	0.478	0.104	(UQ, LQ) at \bar{c}	$0.29 = 0.80 - 0.51$
Interaction model				
Intercept	-5.854	6.694		
Color (0 = dark, 1 = other)	-6.958	7.318		
Width, x (cm)	0.200	0.262	(UQ, LQ) at $c = 0$	$0.13 = 0.43 - 0.30$
Width \times color	0.322	0.286	(UQ, LQ) at $c = 1$	$0.29 = 0.84 - 0.55$

To compare effects of quantitative predictors having different units, it can also be helpful to report standardized coefficients. One approach fits the model to standardized predictors, replacing each x_j by $(x_j - \bar{x}_j)/s_{x_j}$. Then, each regression coefficient represents the effect of a standard deviation change in a predictor, adjusting for the other variables. Equivalently, for each j the standardized coefficient results from multiplying the unstandardized estimate $\hat{\beta}_j$ by s_{x_j} . For example, for fit (5.15) with binary color, the standard deviation of width is 2.109 cm. The standardized estimate for the effect of width for that model is $0.478(2.109) = 1.01$. When we replace width by weight (with standard deviation 0.577 kg) in the model, the unstandardized estimate 1.729 corresponds to the standardized estimate

$1.729(0.577) = 1.00$. The unstandardized estimates 0.478 and 1.729 are quite different, but width and weight (standardized) have similar effects, conditional on whether or not a crab is dark.

Since the standard logistic cdf has standard deviation $\pi/\sqrt{3}$, some software (e.g., PROC LOGISTIC in SAS) defines a standardized estimate by multiplying the unstandardized estimate by $s_{x_j}\sqrt{3}/\pi$. Such a standardized estimate represents the effect on the location of an underlying latent response variable (in standard deviations units) for a standard deviation change in a predictor, adjusting for the other variables. For example, for fit (5.15) with binary color, this standardized estimate for the effect of width is $0.478(2.109)\sqrt{3}/\pi = 0.556$. A standard deviation change in width, conditional on a color, corresponds to a 0.556 standard deviation shift upwards in the distribution of the latent logistic response variable.

5.4.8 Estimating an Average Causal Effect

In many applications the explanatory variable of primary interest specifies two groups to be compared while adjusting for the other explanatory variables in the model. Let $j = 1$ identify this binary group variable, with the groups denoted by $x_1 = 0$ and $x_1 = 1$. For the logistic regression model, an alternative to the log odds ratio β_1 as an effect summary is the estimated *average causal effect*,

$$\frac{1}{n} \sum_i [\hat{\pi}(x_{i1} = 1, x_{i2}, \dots, x_{ip}) - \hat{\pi}(x_{i1} = 0, x_{i2}, \dots, x_{ip})].$$

For each observation i , we find the fitted probability for the given values of x_{i2}, \dots, x_{ip} (1) if that observation were in group 1 and (2) if that observation were in group 0, and average the differences among all n observations. This estimates the difference between the overall proportions of “successes” if all subjects in the study were in group 1 compared with all being in group 0. It is usually not adequate to use a linear probability model (i.e., identity link function) for the full data set, by which such a difference would be constant across subjects, but nonetheless this is a useful summary for cases in which this difference is relatively stable.

We illustrate using [Table 5.6](#) from the randomized study of AZT use and AIDS. In Section 5.4.2 we summarized the effect of AZT use by the estimated conditional odds ratio of $\exp(\beta_1) = 0.487$. Alternatively, from the probability estimates shown in [Table 5.7](#), the difference between those not receiving AZT and those receiving AZT in the estimated proportion developing AIDS symptoms was $0.2654 - 0.1496 = 0.1158$ for whites and $0.2547 - 0.1427 = 0.1120$ for blacks. Weighted by the sample sizes of whites and blacks, the estimated average causal effect is $(220/338)(0.1158) + (118/338)(0.1120) = 0.1145$. In fact, this is similar to the ML estimate of $\beta_1 = 0.1152$ for the corresponding linear probability model.

For categorical predictors, Copas and Eguchi (2010) showed how to obtain a standard error for an estimated average causal effect that applies for the logistic model. They also presented a nonparametric standard error for an estimate that, instead of being model-based, is a weighted average of the differences of the sample proportions at the various levels of the explanatory variables. The main theme of their article, however, was adjusting inferences for the fact that many models may be consistent with the data. The average causal effect is often a relevant measure regardless of the form of the true relationship.

Estimating an average causal effect is natural for experimental studies. It has also received much attention for nonrandomized studies since the fundamental article by Rubin (1974) and later work using methods to adjust for different propensities of a subject to be in one group or the other (e.g., see Section 6.4.11).

5.5 FITTING LOGISTIC REGRESSION MODELS

The mechanics of ML estimation and model fitting for logistic regression are special cases of the GLM fitting results of Section 4.6. With n subjects, we treat the n binary responses as independent. Let $\mathbf{x}_i = (x_{i0}, x_{i1}, \dots, x_{ip})$ denote setting i of the values of p explanatory variables and a coefficient $x_{i0} = 1$ for an intercept term, $i = 1, \dots, N$. When explanatory variables are continuous, a different setting may occur for each subject, in which case $N = n$. This also happens when the data file consists of ungrouped data. The logistic regression model (5.8), treating the intercept α as a regression parameter β_0 for an explanatory variable that always equals 1, is

$$(5.16) \quad \pi(\mathbf{x}_i) = \frac{\exp(\sum_{j=0}^p \beta_j x_{ij})}{1 + \exp(\sum_{j=0}^p \beta_j x_{ij})}.$$

5.5.1 Likelihood Equations for Logistic Regression

When more than one observation occurs at a fixed \mathbf{x}_i value, it is sufficient to record the number of observations n_i and the number of successes. We then let y_i refer to this success count rather than to an individual binary response. Then $\{Y_1, \dots, Y_N\}$ are independent binomials with $E(Y_i) = n_i \pi(\mathbf{x}_i)$, where $n_1 + \dots + n_N = n$. Their joint probability mass function is proportional to the product of N binomial functions,

$$\begin{aligned} & \prod_{i=1}^N \pi(\mathbf{x}_i)^{y_i} [1 - \pi(\mathbf{x}_i)]^{n_i - y_i} \\ &= \left\{ \prod_{i=1}^N \exp \left[\log \left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right)^{y_i} \right] \right\} \left\{ \prod_{i=1}^N [1 - \pi(\mathbf{x}_i)]^{n_i} \right\} \\ &= \left\{ \exp \left[\sum_{i=1}^N y_i \log \frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} \right] \right\} \left\{ \prod_{i=1}^N [1 - \pi(\mathbf{x}_i)]^{n_i} \right\}. \end{aligned}$$

For model (5.16), the i th logit is $\sum_j \beta_j x_{ij}$, so the exponential term here equals $\exp[\sum_i y_i (\sum_j \beta_j x_{ij})] = \exp[\sum_j (\sum_i y_i x_{ij}) \beta_j]$. Also, since $[1 - \pi(\mathbf{x}_i)] = [1 + \exp(\sum_j \beta_j x_{ij})]^{-1}$, the log likelihood equals

$$(5.17) \quad L(\boldsymbol{\beta}) = \sum_j \left(\sum_i y_i x_{ij} \right) \beta_j - \sum_i n_i \log \left[1 + \exp \left(\sum_j \beta_j x_{ij} \right) \right].$$

This depends on the binomial counts only through the sufficient statistics for the model parameters, $\{\sum_i y_i x_{ij}\}$, $j = 0, 1, \dots, p$.

The likelihood equations result from setting $\partial L(\boldsymbol{\beta}) / \partial \boldsymbol{\beta} = \mathbf{0}$. Since

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_i y_i x_{ij} - \sum_i n_i x_{ij} \frac{\exp(\sum_k \beta_k x_{ik})}{1 + \exp(\sum_k \beta_k x_{ik})},$$

the likelihood equations are

$$(5.18) \quad \sum_i y_i x_{ij} - \sum_i n_i \hat{\pi}_i x_{ij} = 0, \quad j = 0, 1, \dots, p,$$

where $\hat{\pi}_i = \exp(\sum_k \hat{\beta}_k x_{ik}) / [1 + \exp(\sum_k \hat{\beta}_k x_{ik})]$ is the ML estimate of $\pi(\mathbf{x}_i)$. We observed these equations as a special case of those for binomial GLMs in (4.28) (but there y_i is the proportion of successes). The equations are nonlinear and require iterative solution.

Let \mathbf{X} denote the matrix of values of $\{x_{ij}\}$, with N rows for the binomial observations and a column for each parameter. The likelihood equations (5.18) have form

$$(5.19) \quad \mathbf{X}^T \mathbf{y} = \mathbf{X}^T \hat{\boldsymbol{\mu}},$$

where $\hat{\boldsymbol{\mu}}_i = n_i \hat{\pi}_i$. This equation illustrates the fundamental result for GLMs with canonical link, shown in equation (4.51), that the likelihood equations equate the sufficient statistics to their expected values.

5.5.2 Asymptotic Covariance Matrix of Parameter Estimators

The ML estimators $\hat{\beta}$ have a large-sample normal distribution with covariance matrix equal to the inverse of the information matrix. The observed information matrix has elements

$$(5.20) \quad -\frac{\partial^2 L(\beta)}{\partial \beta_a \partial \beta_b} = \sum_i \frac{x_{ia}x_{ib}n_i \exp(\sum_j \beta_j x_{ij})}{[1 + \exp(\sum_j \beta_j x_{ij})]^2} = \sum_i x_{ia}x_{ib}n_i \pi_i(1 - \pi_i).$$

This is not a function of $\{y_i\}$, so the observed and expected information are identical. This happens for all GLMs that use canonical links (Section 4.6.5).

The estimated covariance matrix is the inverse of the matrix having elements (5.20), substituting $\hat{\beta}$. This has form

$$(5.21) \quad \widehat{\text{cov}}(\hat{\beta}) = \{X^T \text{Diag}[n_i \hat{\pi}_i(1 - \hat{\pi}_i)] X\}^{-1},$$

where $\text{Diag}[n_i \hat{\pi}_i(1 - \hat{\pi}_i)]$ denotes the $N \times N$ diagonal matrix having $\{n_i \hat{\pi}_i(1 - \hat{\pi}_i)\}$ on the main diagonal. This is the special case of the GLM covariance matrix (4.31) with estimated diagonal weight matrix \hat{W} having elements $\hat{w}_i = n_i \hat{\pi}_i(1 - \hat{\pi}_i)$. The square roots of the main diagonal elements of (5.21) are estimated standard errors of $\hat{\beta}$.

5.5.3 Distribution of Probability Estimators

Using $\widehat{\text{cov}}(\hat{\beta})$, we can conduct Wald inference about β and related effects such as odds ratios. We can also construct confidence intervals for response probabilities $\pi(x)$ at particular settings $x^T = (x_0, x_1, \dots, x_p)$.

The estimated variance of $\text{logit}[\hat{\pi}(x)] = x^T \hat{\beta}$ is $x^T \widehat{\text{cov}}(\hat{\beta}) x$. For large samples, $\text{logit}[\hat{\pi}(x)] \pm z_{\alpha/2} \sqrt{x^T \widehat{\text{cov}}(\hat{\beta}) x}$ is a confidence interval for the true logit. The endpoints invert to a corresponding interval for $\pi(x)$ using the transform $\pi = \exp(\text{logit})/[1 + \exp(\text{logit})]$.

5.5.4 Newton–Raphson Method Applied to Logistic Regression

Section 4.6.1 introduced the Newton–Raphson iterative method, which applies in a straightforward manner to logistic regression. Let

$$\begin{aligned} u_j^{(t)} &= \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} \Big|_{\boldsymbol{\beta}^{(t)}} = \sum_i (y_i - n_i \pi_i^{(t)}) x_{ij}, \\ h_{ab}^{(t)} &= \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_a \partial \beta_b} \Big|_{\boldsymbol{\beta}^{(t)}} = - \sum_i x_{ia} x_{ib} n_i \pi_i^{(t)} (1 - \pi_i^{(t)}). \end{aligned}$$

Here, $\boldsymbol{\pi}^{(t)}$, approximation t for $\hat{\boldsymbol{\pi}}$ is obtained from $\boldsymbol{\beta}^{(t)}$ through

$$(5.22) \quad \pi_i^{(t)} = \frac{\exp\left(\sum_{j=1}^p \beta_j^{(t)} x_{ij}\right)}{1 + \exp\left(\sum_{j=1}^p \beta_j^{(t)} x_{ij}\right)}.$$

We use $\mathbf{u}^{(t)}$ and $\mathbf{H}^{(t)}$ with formula (4.45) to obtain the next value $\boldsymbol{\beta}^{(t+1)}$, which in this context is

$$(5.23) \quad \boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \{\mathbf{X}^T \text{Diag}[n_i \pi_i^{(t)} (1 - \pi_i^{(t)})] \mathbf{X}\}^{-1} \mathbf{X}^T (\mathbf{y} - \boldsymbol{\mu}^{(t)}),$$

where $\boldsymbol{\mu}_i^{(t)} = n_i \boldsymbol{\pi}_i^{(t)}$. This is used to obtain $\boldsymbol{\pi}^{(t+1)}$, and so forth.

With an initial guess $\boldsymbol{\beta}^{(0)}$, (5.22) yields $\boldsymbol{\pi}^{(0)}$, and for $t > 0$ the iterations proceed as just described using (5.23) and (5.22). In the limit, $\boldsymbol{\pi}^{(t)}$ and $\boldsymbol{\beta}^{(t)}$ converge to the ML estimates $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\beta}}$ (Walker and Duncan 1967). The $\mathbf{H}^{(t)}$ matrices converge to $\hat{\mathbf{H}} = -\mathbf{X}^T \text{Diag}[n_i \hat{\pi}_i (1 - \hat{\pi}_i)] \mathbf{X}$. By (5.21) the estimated asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is a by-product of the Newton–Raphson method, namely $-\hat{\mathbf{H}}^{-1}$.

From the argument in Section 4.6.4, $\boldsymbol{\beta}^{(t+1)}$ has the iterative reweighted least-squares form $(\mathbf{X}^T \mathbf{V}_t^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}_t^{-1} \mathbf{z}^{(t)}$, where $\mathbf{z}^{(t)}$ has elements

$$(5.24) \quad z_i^{(t)} = \log \frac{\pi_i^{(t)}}{1 - \pi_i^{(t)}} + \frac{y_i - n_i \pi_i^{(t)}}{n_i \pi_i^{(t)} (1 - \pi_i^{(t)})},$$

and where \mathbf{V}_t is a diagonal matrix with elements $\{1/n_i \pi_i^{(t)} (1 - \pi_i^{(t)})\}$. In this expression, $\mathbf{z}^{(t)}$ is the linearized form of the logit link function for the sample data, evaluated at $\boldsymbol{\pi}^{(t)}$ [see (4.49)]. From Section 3.1.6 the elements of \mathbf{V}_t are estimated asymptotic variances of the sample logits. The ML estimate is the limit of a sequence of weighted least-squares estimates, where the weight matrix changes at each cycle.

The log likelihood is concave, so there is no danger of iterative methods converging to a local maximum. However, in some cases at least one estimate may be infinite, as discussed in Section 6.5.

NOTES

Section 5.1: Interpreting Parameters in Logistic Regression

5.1 Logistic books: Books focusing on logistic regression include Collett (2003), Cox and Snell (1989), and Hosmer and Lemeshow (2000).

5.2 Bias reduction: Haldane (1956) recommended adding $\frac{1}{2}$ to each count in estimating a logit. With this modification, the bias is on the order of only $1/n_i^2$, for large n_i . See also Firth (1993a), Gait and Zweifel (1967), and Exercise 16.8. For bias reduction in logistic regression and GLMs, see Cordeiro and McCullagh (1991) and Firth (1993a).

5.3 LD₅₀: Paige et al. (2011) summarized confidence intervals for LD₅₀ and proposed small-sample intervals using saddlepoint approximations.

5.4 Retrospective logistic: For discussion of logistic regression with retrospective studies, see Anderson (1972), Breslow (1996), Breslow and Day (1980, p. 203), Breslow and Powers (1978), Carroll et al. (1995), Farewell (1979), Ghosh and Mukherjee (2010), Mantel (1973), Neuhaus and Jewell (1990b), Piegorsch et al. (1994), Prentice (1976a), Prentice and Pyke (1979), Roeder et al. (1996), and Umbach and Weinberg (1997). Scott and Wild (2001) considered case-control studies with complex sampling designs, and Bhadra et al. (2012) incorporated longitudinal information on exposure history. Qin and Liang (2011) considered a mixture model to handle situations in which some controls are contaminated. See Section 7.2.3 for Bayesian literature.

5.5 Design: Khuri et al. (2006) reviewed articles about design problems for binary response experiments. Issues include choosing settings for a predictor to optimize a criterion for estimating parameter values, and estimating the setting at which the response probability equals some fixed value. The nonconstant variance makes this challenging. Zocchi and Atkinson (1999) considered multinomial logistic models.

Section 5.2: Inference for Logistic Regression

5.6 Fitting/checking: Albert and Anderson (1984), Berkson (1944, 1951, 1953, 1955), Cox (1958a), Hodges (1958), and Walker and Duncan (1967) discussed ML estimation for logistic regression, although Berkson argued for the computationally simpler minimum logit chi-squared. For adjustments with complex sample surveys, see Hosmer and Lemeshow (2000, Sec. 6.4) and LaVange et al. (2001). Grouping values to check model fit extends to any GLM (Pregibon 1982). Hosmer et al. (1997) compared various ways to do this. Presnell and Boos (2004) proposed a general likelihood-based method for detecting model misspecification. See also Capanu and Presnell (2008).

Section 5.3: Logistic Models with Categorical Predictors

5.7 Trend tests: Extensions of the trend test include handling of correlated binary data by Corcoran et al. (2001) and stratified $I \times J$ tables by Mantel (1963). Williams (2005) surveyed trend tests for proportions and counts.

Section 5.4: Multiple Logistic Regression

5.8 Standardizing: Menard (2004) discussed several approaches to standardizing logistic regression coefficients. He noted that merely standardizing predictors, as was done in Section 5.4.7, is adequate for comparing influences of predictors.

5.9 Quasi-variances: For multipredictor models such as (5.12), tables that contain factor-level estimates (β^x) and their SE values but not their covariance matrix permit comparison of each category to the baseline (having estimate 0) but not to other categories. Firth and De Menezes

(2004) showed how to construct *quasi-variances* $\{q_k\}$ such that the SE of $\hat{\beta}_a^x - \hat{\beta}_b^x$ is approximately $\sqrt{q_a^2 + q_b^2}$.

EXERCISES

Applications

5.1 An article about the contributions of star players in the National Basketball Association (by M. L. Jones and R. J. Parker, *Chance*: **23**, 39–45, 2010) reported prediction equations for the probability π of a win in a game for a player, using as predictors $ortg$ = player's offensive rating in the game, which is the number of points produced per hundred possessions, $drtg$ = player's defensive rating in the game, which is the number of points allowed per hundred possessions (the lower the better), and $home$, which indicates whether the game was played at home (1 = yes, 0 = no). For LeBron James using data from the 2008-2009 season,

$$\text{logit}(\hat{\pi}) = 1.379 + 0.119(ortg) - 0.139(drtg) + 3.393(home).$$

- a. Over the season, James's quartiles (lower, median, upper) were (108.7, 123.2, 136.1) for $ortg$ and (91.7, 99.5, 107.7) for $drtg$. Summarize the $ortg$ effect for James by comparing $\hat{\pi}$ at its upper and lower quartiles. Do this at the median level of $drtg$, separately for home and away games. Repeat for the $drtg$ effect, and compare.
- b. Summarize the $home$ effect by (i) comparing $\hat{\pi}$ for home and away games, at the median levels of $ortg$ and $drtg$, (ii) interpreting its coefficient in the fitted logistic equation.

5.2 For a study using logistic regression to determine characteristics associated with remission in cancer patients, [Table 5.11](#) shows the most important explanatory variable, a labeling index (LI) that measures proliferative activity of cells after a patient receives an injection of tritiated thymidine. It represents the percentage of cells that are “labeled.” The response measured whether the patient achieved remission. Software reports [Table 5.12](#) for a logistic regression model using LI to estimate $\pi = P(\text{remission})$.

[Table 5.11](#) Data for Exercise 5.2 on Cancer Remission

LI	Number of Cases		Number of Remissions		LI	Number of Cases		Number of Remissions	
	Number of Cases	Number of Remissions	Number of Cases	Number of Remissions		Number of Cases	Number of Remissions	Number of Cases	Number of Remissions
8	2	0	18	1	1	28	1	1	1
10	2	0	20	3	2	32	1	0	0
12	3	0	22	2	1	34	1	1	1
14	3	0	24	1	0	38	3	2	2
16	3	0	26	1	1				

Source: Data reprinted with permission from E. T. Lee, *Comput. Prog. Biomed.* **4**: 80–92, 1974.

[Table 5.12](#) Software Output (Based on SAS) for Exercise 5.2

Parameter	Estimate	Standard Error	Intercept	Intercept and Covariates
			Only	Covariates
-2 Log L	34.372		34.372	26.073
Intercept	-3.7771	1.3786		7.5064
li	0.1449	0.0593		5.9594
				0.0061
				0.0146
Odds Ratio Estimates				
Effect	Point Estimate	95% Wald Confidence Limits		
li	1.156	1.029	1.298	
Estimated Covariance Matrix				
Variable	Intercept	li		
Intercept	1.900616	-0.07653		
li	-0.07653	0.003521		
Obs	li	remiss	n	pihat
1	8	0	2	0.06797
2	10	0	2	0.08879
				lower upper
1				0.01121 0.31925
2				0.01809 0.34010

- a. Show how software obtained $\hat{\pi} = 0.068$ when LI = 8.
- b. Show that $\pi = 0.50$ when LI = 26.0.
- c. Show that the rate of change in $\hat{\pi}$ is 0.009 when LI = 8 and 0.036 when LI = 26.

- d. The lower quartile and upper quartile for LI are 14 and 28. Show that $\hat{\pi}$ increases by 0.42, from 0.15 to 0.57, between those values.
- e. For a unit increase in LI, show that the estimated odds of remission multiply by 1.16.
- f. Explain how to obtain the confidence interval reported for the odds ratio. Interpret.
- g. Construct a Wald test for the effect. Interpret.
- h. Conduct a likelihood-ratio test for the effect, showing how to construct the test statistic using the $-2 \log L$ values reported.
- i. Show how software obtained the confidence interval for π reported at LI = 8. [Hint: Use the reported covariance matrix.]

5.3 The text website has a data file (created from data at www.basketball-reference.com) showing, for each game in the 2010–2011 season of the National Basketball Association in which Rajon Rondo of the Boston Celtics played, x = the number of assists he recorded and y = whether the Celtics won (1 = yes). Using software, (a) show that the logistic model fitted to these data gives $\text{logit}[\hat{p}(Y=1)] = -2.235 + 0.294x$; (b) show that $\hat{p}(Y=1)$ increases from 0.21 to 0.99 over the observed range of x from 3 to 24; and (c) construct a significance test and confidence interval about the effect in the conceptual population that these games represent.

5.4 [Table 5.13](#) summarizes logistic regression results from a study¹ of how family transitions relate to first home purchase by young married households. The response variable is whether the subject owns a home (1 = yes, 0 = no).

Table 5.13 Results of Logistic Regression for Probability of Home Ownership

Variable	Estimate	Std. Error
Intercept	-2.870	—
Husband earnings (\$10,000)	0.569	0.088
Wife earnings (\$10,000)	0.306	0.140
Number of years married	-0.039	0.042
Married in 2 years (1 = yes)	0.224	0.304
Working wife in 2 years (1 = yes)	0.373	0.283
Number of children	0.220	0.101
Add child in 2 years (1 = yes)	0.271	0.140
Head's education (no. years)	0.027	0.032
Parents' home ownership (1 = yes)	0.387	0.176

- a. Interpret the effects that seem to be significant.
- b. Fill in the blanks: Adjusting for the other explanatory variables, each additional child had the effect of multiplying the estimated odds of owning a home by ____; that is, the estimated odds increase by ____%. A \$10,000 increase in earnings had the effect of multiplying the estimated odds of owning a home by ____ if the earnings add to husband's income and by ____ for wife's income.

5.5 Consider the fit of model [\(5.2\)](#) for the horseshoe crabs using x = width.

- a. Show that (i) at the mean width (26.3), the estimated odds of a satellite equal 2.07; (ii) at $x = 27.3$, the estimated odds equal 3.40; and (iii) since $\exp(0) = 1.64$, $3.40 = (1.64)2.07$, and the odds increase by 64%.
- b. Based on the 95% confidence interval for β , show that for x near where $\pi = 0.50$, the rate of increase in the probability of a satellite per 1-cm increase in x falls between about 0.07 and 0.17.

5.6 For the 23 space shuttle flights before the *Challenger* mission disaster in 1986, [Table 5.14](#) shows the temperature at the time of the flight and whether at least one primary O-ring suffered thermal distress.

Table 5.14 Data for Exercise 5.6 on Challenger Space-Shuttle Disaster^a

Ft	Temp	TD												
1	66	0	2	70	1	3	69	0	4	68	0	5	67	0
6	72	0	7	73	0	8	70	0	9	57	1	10	63	1
11	70	1	12	78	0	13	67	0	14	53	1	15	67	0
16	75	0	17	70	0	18	81	0	19	76	0	20	79	0
21	75	1	22	76	0	23	58	1						

^aFt, flight number; Temp, temperature ($^{\circ}$ F); TD, thermal distress (1, yes; 0, no).

Source: Data based on Table 1 in *J. Am. Statist. Assoc.* 84: 945–957, 1989, by S. R. Dalal, E. B. Fowlkes, and B. Hoadley. Reprinted with permission from *J. Am. Statist. Assoc.*

- a. Use logistic regression to model the effect of temperature on the probability of thermal distress. Plot a figure of the fitted model, and interpret.
- b. Estimate the probability of thermal distress at 31°F, the temperature at the place and time of the *Challenger* flight.
- c. Construct a confidence interval for the effect of temperature on the odds of thermal distress, and test the statistical significance of the effect.

5.7 Refer to [Table 4.2](#). Using scores (0, 2, 4, 5) for snoring, fit the logistic regression model. Interpret using fitted probabilities, linear approximations, and effects on the odds. Analyze the goodness of fit.

5.8 Hastie and Tibshirani (1990, p. 282) described a study to determine risk factors for kyphosis, severe forward flexion of the spine following corrective spinal surgery. The age in months at the time of the operation for the 18 subjects for whom kyphosis was present were 12, 15, 42, 52, 59, 73, 82, 91, 96, 105, 114, 120, 121, 128, 130, 139, 139, 157 and for 22 of the subjects for whom kyphosis was absent were 1, 1, 2, 8, 11, 18, 22, 31, 37, 61, 72, 81, 97, 112, 118, 127, 131, 140, 151, 159, 177, 206.

- a. Fit a logistic regression model using age as a predictor of whether kyphosis is present. Test whether age has a significant effect.
- b. Plot the data. Note the difference in dispersion on age at the two levels of kyphosis. Fit the model $\text{logit}[\pi(x)] = \alpha + \beta_1 x + \beta_2 x^2$. Test the significance of the squared age term, plot the fit, and interpret. (See also Exercise 5.30 and Section 7.4.3.)

5.9 For [Table 5.5](#) on treating leprosy, the Pearson test of independence has $X^2(I) = 6.88$ ($P = 0.14$). For equally spaced scores, the Cochran–Armitage trend test has $z^2 = 6.67$ ($P = 0.01$). Interpret, and explain why the P -values differ so. Analyze the data, using a linear logit model. Test independence using the Wald and likelihood-ratio tests, and compare results to the Cochran–Armitage test. Check the fit of the model, and interpret.

5.10 Refer to [Table 5.3](#) on infant malformation and alcohol consumption.

- a. Repeat the trend test of Section 5.3.5 after deleting the single case in the last row. Comment on that observation's influence.
- b. Repeat the trend test using alcohol consumption scores (1, 2, 3, 4, 5) instead of (0.0, 0.5, 1.5, 4.0, 7.0). Compare results, noting the potential sensitivity to the choice of scores for highly unbalanced data.

5.11 A study used the 1998 Behavioral Risk Factors Social Survey to consider factors associated with women's use of oral contraceptives in the United States. [Table 5.15](#) summarizes effects for a logistic regression model for the probability of using oral contraceptives. Each predictor uses an indicator variable, and the table lists the category having indicator outcome 1. Interpret effects. Construct and interpret a confidence interval for the conditional odds ratio between contraceptive use and education.

Table 5.15 Data for Exercise 5.11 on Oral Contraceptive Use

Variable	Coding = 1 if:	Estimate	SE
Age	35 or younger	−1.320	0.087
Race	White	0.622	0.098
Education	≥1 year college	0.501	0.077
Marital status	Married	−0.460	0.073

Source: Data courtesy of Debbie Wilson, College of Pharmacy, University of Florida.

5.12 For the horseshoe crab data, available at www.stat.ufl.edu/~aa/cda/cda.html, fit a logistic regression model for the probability of a satellite, using color alone as the predictor.

- Treat color as nominal. Explain why this model is saturated. Express its parameter estimates in terms of the sample logits for each color.
- Conduct a likelihood-ratio test that color has no effect.
- Fit a model that treats color as quantitative. Interpret the fit, and test that color has no effect.
- Test the goodness of fit of the model in part (c). Interpret.

5.13 For model (5.15) with binary color c and width x , (a) describe the effect of width by finding the estimated probabilities of a satellite at its lower and upper quartiles, separately for $c = 1$ and $c = 0$, and (b) describe the effect of color by its average causal effect.

5.14 Refer to the prediction equation $\text{logit}(\pi) = -10.071 - 0.509c + 0.458x$ for model (5.14) using quantitative color and width. The means and standard deviations are $\bar{c} = 2.44$ and $s = 0.80$ for color, and $\bar{x} = 26.30$ and $s = 2.11$ for width. For standardized predictors [e.g., $x = (\text{width} - 26.30)/2.11$], explain why the estimated coefficients of c and x equal -0.41 and 0.97 . Interpret these by comparing the partial effects of a standard deviation increase in each predictor on the odds. Describe the color effect by estimating the change in π between the first and last color categories at the sample mean width.

5.15 For [Table 2.6](#), we fitted a logistic model, treating death penalty as the response (1 = yes) and defendant's race (1 = white) and victims' race (1 = white) as indicator predictors. [Table 5.16](#) shows results.

Table 5.16 Software Output (Based on SAS) for Exercise 5.15 on the Death Penalty

Criteria For Assessing Goodness Of Fit					
Criterion	DF	Value			
Deviance	1	0.3798			
Pearson Chi-Square	1	0.1978			
Log Likelihood		-209.4783			
Standardized Parameter Estimates					
Parameter	Estimate	Error	95% Conf Limits	Likelihood Ratio	Chi-Square
Intercept	-3.5961	0.5069	-4.7754	-2.7349	50.33
def	-0.8678	0.3671	-1.5633	-0.1140	5.59
vic	2.4044	0.6006	1.3068	3.7175	16.03
LR Statistics					
Source	DF	Chi-Square	Pr > Chisq		
def	1	5.01	0.0251		
vic	1	20.35	<.0001		

- Interpret parameter estimates. Which group is most likely to have the yes response? Find the estimated probability in that case.
- Interpret 95% confidence intervals for conditional odds ratios.
- Test the effect of defendant's race, controlling for victims' race, using a (i) Wald test and (ii) likelihood-ratio test. Interpret.
- Test the goodness of fit of the model. Interpret.

5.16 Model the effects of victim's race and defendant's race for [Table 2.12](#). Interpret.

5.17 In a 2011 article in *North Carolina Law Review*, M. Radelet and G. Pierce reported a logistic prediction equation for death penalty verdicts in North Carolina. Let Y denote whether a subject convicted of murder received the death penalty (1 = yes), for defendant's race h ($h = 1$, black; $h = 2$, white), victim's race i ($i = 1$, black; $i = 2$, white), and number of additional factors j ($j = 0, 1, 2$). For the model

$$\text{logit}[P(Y = 1)] = \alpha + \beta_h^D + \beta_i^V + \beta_j^F$$

they reported $\hat{\alpha} = -5.26$, $\hat{\beta}_1^D = 0.00$, $\hat{\beta}_2^D = 0.17$, $\hat{\beta}_1^V = 0.00$, $\hat{\beta}_2^V = 0.91$, $\hat{\beta}_0^F = 0.00$, $\hat{\beta}_1^F = 2.02$, $\hat{\beta}_2^F = 3.98$.

- Estimate the probability of receiving the death penalty for the group most likely to receive

it.

- b. If, instead, parameters used constraints $\beta_2^D = \beta_2^V = \beta_2^F = 0$, report the estimates.
- c. If, instead, parameters used constraints $\sum_h \beta_h^D = \sum_i \beta_i^V = \sum_j \beta_j^F = 0$ report the estimates.

5.18 In a study designed to evaluate whether an educational program makes sexually active adolescents more likely to obtain condoms, adolescents were randomly assigned to two experimental groups. The educational program, involving a lecture and videotape about transmission of HIV, was provided to one group but not the other. [Table 5.17](#) summarizes results of a logistic regression model for factors observed to influence teenagers to obtain condoms.

[Table 5.17](#) Data for Exercise 5.18 on Obtaining Condoms

Variable	Odds Ratio	95% Confidence Interval
Group (education vs. none)	4.04	(1.17, 13.9)
Gender (males vs. females)	1.38	(1.23, 12.88)
SES (high vs. low)	5.82	(1.87, 18.28)
Lifetime number of partners	3.22	(1.08, 11.31)

Source: V. I. Rickert et al., *Clin. Pediatr.* **31**: 205–210, 1992.

- a. Find the parameter estimates for the fitted model, using (1, 0) indicator variables for the first three predictors. Based on the corresponding confidence interval for the log odds ratio, determine the standard error for the group effect.
- b. Explain why either the estimate of 1.38 for the odds ratio for gender or the corresponding confidence interval is incorrect. Show that if the reported interval is correct, 1.38 is actually the *log* odds ratio, and the estimated odds ratio equals 3.98.

5.19 [Table 5.18](#) shows estimated effects for a logistic regression model with squamous cell esophageal cancer ($Y=1$, yes; $Y=0$, no) as the response. Smoking status (S) equals 1 for at least one pack per day and 0 otherwise, alcohol consumption (A) equals the average number of alcoholic drinks consumed per day, and race (R) equals 1 for blacks and 0 for whites. To describe the $R \times S$ interaction, construct the prediction equation when $R = 1$ and again when $R = 0$. Find the fitted YS conditional odds ratio for each case. Similarly, construct the prediction equation when $S = 1$ and again when $S = 0$. Find the fitted YR conditional odds ratios. Note that for each association, the coefficient of $R \times S$ is the difference between the log odds ratios at the two fixed levels for the other variable. Explain why the coefficient of S represents the log odds ratio between Y and S for whites. To what hypotheses do the *P*-values for R and S refer?

[Table 5.18](#) Data for Exercise 5.19 on Esophageal Cancer

Variable	Effect	<i>P</i> -value
Intercept	-7.00	< 0.01
Alcohol use (A)	0.10	0.03
Smoking (S)	1.20	< 0.01
Race (R)	0.30	0.02
Race \times smoking ($R \times S$)	0.20	0.04

5.20 A survey of high school students on Y = whether the subject has driven a motor vehicle after consuming a substantial amount of alcohol (1 = yes), s = sex (1 = female), r = race (1 = black; 0 = white), and g = grade ($g_1 = 1$, grade 9; $g_2 = 1$, grade 10; $g_3 = 1$, grade 11; $g_1 = g_2 = g_3 = 0$, grade 12) has prediction equation

$$\begin{aligned}\text{logit}[\hat{P}(Y = 1)] &= -0.88 - 0.40s - 0.72r - 2.22g_1 - 1.43g_2 - 0.58g_3 \\ &\quad + 0.74(r \times g_1) + 0.38(r \times g_2) + 0.01(r \times g_3).\end{aligned}$$

Carefully interpret effects. Explain the interaction by describing the race effect at each grade and the grade effect for each race.

5.21 The Gallup Poll reported in March 2010 that the percentage believing that news reports exaggerate the seriousness of global warming is 66% for Republicans and 22% for Democrats. By contrast, in 1998 the corresponding percentages were 34% and 23%. Considered as results

for a three-way table cross-classifying opinion by political party and year, do these data seem to display interaction? In what sense?

5.22 A table at the text website refers to a sample of subjects randomly selected for an Italian study on the relation between income and whether one possesses a travel credit card. At each level of annual income in millions of lira (the Italian currency at the time of the study), the table indicates the number of subjects sampled and the number possessing at least one travel credit card. Analyze these data.

5.23 A research article in the *British Medical Journal* (by C. de Oliveira et al., 2010, vol. 340) showed results from the Scottish Health Survey, indicating that over a period of about 8 years, cardiovascular disease events occurred for 308 of 8481 subjects who reported brushing their teeth at least twice a day, for 188 of 2850 subjects who reported brushing once a day, and for 59 of 538 subjects who reported brushing less than once a day. Analyze these data.

5.24 Are people with more social ties less likely to get colds? Use logistic models to analyze the $2 \times 2 \times 2 \times 2$ contingency table on p. 1943 of the article by S. Cohen et al., *J. Am. Med. Assoc.* **277** (24).

Theory and Methods

5.25 For logistic regression model (5.1), show that $\partial\pi(x)/\partial x = \beta\pi(x)[1 - \pi(x)]$.

5.26 For logistic model (5.1), when $\pi(x)$ is small, explain why you can interpret $\exp(\beta)$ approximately as $\pi(x+1)/\pi(x)$.

5.27 Prove that the logistic regression curve (5.1) has the steepest slope where $\pi(x) = \frac{1}{2}$. Generalize to model (5.8).

5.28 The calibration problem is that of estimating x at which $\pi(x) = \pi_0$ for some fixed π_0 such as 0.50. For the linear logit model, argue that a confidence interval is the set of x values for which

$$|\hat{\alpha} + \hat{\beta}x - \text{logit}(\pi_0)|/[\text{var}(\hat{\alpha}) + x^2 \text{ var}(\hat{\beta}) + 2x \text{ cov}(\hat{\alpha}, \hat{\beta})]^{1/2} < z_{\alpha/2}.$$

An alternative approach inverts a likelihood-ratio test.

5.29 A study for several professional sports of the effect of a player's draft position d ($d = 1, 2, 3, \dots$) of selection from the pool of potential players in a given year on the probability π of eventually being named an all star used the model $\text{logit}(\pi) = \alpha + \beta \log d$ (S. M. Berry, *Chance*, **14**(2): 53–57, 2001).

a. Show that $\pi/(1 - \pi) = e^\alpha d^\beta$. Show that $e^\alpha = \text{odds for the first draft pick}$.

b. In the United States, Berry reported $\hat{\alpha} = 2.3$ and $\hat{\beta} = -1.1$ for pro basketball and $\hat{\alpha} = 0.7$ and $\hat{\beta} = -0.6$ for pro baseball. This suggests that in basketball a first draft pick is more crucial and picks with high d are relatively less likely to be all-stars. Explain why.

5.30 For the population having $Y=j$, supposed has a $N(\mu_j, \sigma^2)$ distribution, $j = 0, 1$.

a. Using Bayes' theorem, show that $P(Y=1|x)$ satisfies the logistic regression model with $\beta = (\mu_1 - \mu_0)/\sigma^2$.

b. Suppose that $(X|Y=j)$ is $N(\mu_j, \sigma_j^2)$ with $\sigma_0 \neq \sigma_1$. Show that the logistic model holds with a quadratic term (Anderson 1975). [Exercise 5.8 showed that a quadratic term is helpful when x values have quite different dispersion at $y=0$ and $y=1$. This result also suggests that to test equality of means of normal distributions when the variances differ, we can fit a quadratic logistic regression with the two groups as the response and test the linear and quadratic terms together; see O'Brien (1988).]

c. Suppose that $(X|Y=j)$ has an exponential family density $f(x; \theta_j) = a(\theta_j)b(x)\exp[xQ(\theta_j)]$. Show that $P(Y=1|x)$ satisfies the logistic model, with effect of x equal to $[Q(\theta_1) - Q(\theta_0)]$.

d. For multiple predictors, suppose that $(X|Y=j)$ has a multivariate $N(\mu_j, \Sigma)$ distribution, $j = 0, 1$. Show that $P(Y=1|x)$ satisfies logistic regression with effect parameters $\Sigma^{-1}(\mu_1 - \mu_0)$ (Cornfield 1962, Warner 1963).

5.31 Suppose that $\pi(x) = F(x)$ for some strictly increasing cdf F . Explain why a monotone

transformation of x exists such that the logistic regression model holds. Generalize to alternative link functions.

5.32 For an $I \times 2$ contingency table, consider logistic model (5.4).

- a. Given $\{\pi_i > 0\}$, show how to find $\{\beta_i\}$ satisfying $\beta_I = 0$.
- b. Prove that $\beta_1 = \beta_2 = \dots = \beta_I$ is the independence model. Find its likelihood equation, and show that $\hat{\alpha} = \text{logit}[(\sum_i y_i)/(\sum_i n_i)]$.

5.33 For a multinomial distribution, let $\gamma = \sum_i b_i \pi_i$, and suppose that $\pi_i = f_i(\theta) > 0$, $i = 1, \dots, I$. For sample proportions $\{p_i\}$, let $S = \sum_i b_i p_i$. Let $T = \sum_i b_i \hat{\pi}_i$, where $\hat{\pi}_i = f_i(\hat{\theta})$, for the ML estimator $\hat{\theta}$ of θ .

- a. Show that $\text{var}(S) = [\sum_i b_i^2 \pi_i - (\sum_i b_i \pi_i)^2]/n$.
- b. Using the delta method, show $\text{var}(T) \approx [\text{var}(\hat{\theta})][\sum_i b_i f'_i(\theta)]^2$.
- c. By computing the information for $L(\theta) = \sum_i n_i \log[f_i(\theta)]$, show that $\text{var}(\hat{\theta})$ is approximately $[n \sum_i (f'_i(\theta))^2/f_i(\theta)]^{-1}$.
- d. Asymptotically, show that $\text{var}[\sqrt{n}(T - \gamma)] \leq \text{var}[\sqrt{n}(S - \gamma)]$. [Hint: Show that $\text{var}(T)/\text{var}(S)$ is a squared correlation between two random variables, where with probability π_i the first equals b_i and the second equals $f'_i(\theta)/f_i(\theta)$.]

5.34 Construct the log-likelihood function for the model $\text{logit}[\pi(x)] = \alpha + \beta x$ with independent binomial outcomes of y_0 successes in n_0 trials at $x = 0$ and y_1 successes in n_1 trials at $x = 1$. Derive the likelihood equations, and show that $\hat{\beta}$ is the sample log odds ratio.

5.35 A study has n_i independent binary observations $\{y_{i1}, \dots, y_{in_i}\}$ when $X = x_i$, $i = 1, \dots, N$, with $n = \sum_i n_i$. Consider the model $\text{logit}(\pi_i) = \alpha + \beta x_i$, where $\pi_i = P(Y_{ij} = 1)$.

- a. Show that the kernel of the likelihood function is the same treating the data as n Bernoulli observations or N binomial observations.
- b. For the saturated model, explain why the likelihood function is different for these two data forms. [Hint: The number of parameters differs.] Hence, the deviance reported by software depends on the form of data entry.
- c. Explain why the difference between deviances for two unsaturated models does not depend on the form of data entry.
- d. Suppose that each $n_i = 1$. Show that the deviance depends on $\hat{\pi}_i$ but not y_i . Hence, it is not useful for checking model fit (see also Exercise 4.18).

5.36 Suppose that Y has a $\text{bin}(n, \pi)$ distribution. For the model, $\text{logit}(\pi) = \alpha$, consider testing $H_0: \alpha = 0$ (i.e., $\pi = 0.50$). Let $\hat{\pi} = y/n$.

- a. Compare the estimated SE for the Wald test and the SE using the null value 0.50 for π , for two possible denominators in the test statistic $[\text{logit}(\hat{\pi})/SE]^2$. Show that the ratio of the Wald statistic to the statistic with null SE equals $4\hat{\pi}(1 - \hat{\pi})$. What is the implication about performance of the Wald test if $|\alpha|$ is large and $\hat{\pi}$ tends to be near 0 or 1?
- b. How does the comparison of tests change with the scale $[(\hat{\pi} - 0.5)/SE]^2$, where SE is now the estimated or null SE of $\hat{\pi}$? [Analogous results apply for inference about the Poisson mean versus the log mean; see also Mantel (1987a) and Section 5.2.6.]

5.37 Find the likelihood equations for model (5.10) with two binary predictors. Show that they imply that the fitted values and the sample counts are identical in the marginal two-way tables.

5.38 Consider the likelihood equations (5.18) for a logistic regression model. Using the equation resulting from the intercept parameter, show that the overall sample proportion of successes equals the sample mean of the fitted success probabilities.

5.39 Consider the linear logit model (5.5) for an $I \times 2$ table, with y_i a $\text{bin}(n_i, \pi_i)$ variate.

- a. Show that the log likelihood is

$$L(\boldsymbol{\beta}) = \sum_{i=1}^I y_i(\alpha + \beta x_i) - \sum_{i=1}^I n_i \log[1 + \exp(\alpha + \beta x_i)].$$

b. Show that the sufficient statistic for β is $\sum_i y_i x_i$, and explain why this is essentially the variable utilized in the Cochran–Armitage test. (That test is a score test of $H_0: \beta = 0$.)

c. Letting $S = \sum_i y_i$, show that the likelihood equations are

$$S = \sum_i n_i \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)},$$

$$\sum_i y_i x_i = \sum_i n_i x_i \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}.$$

d. Let $\{\hat{\mu}_i = n_i \hat{\pi}_i\}$. Explain why $\sum_i \hat{\mu}_i = \sum_i y_i$ and

$$\sum_i x_i \frac{y_i}{S} = \sum_i x_i \frac{\hat{\mu}_i}{\sum_a \hat{\mu}_a}.$$

Explain why this implies that the mean score on x across the rows in the first column is the same for the model fit as for the observed data. (They are also identical for the second column.)

5.40 Let Y_i be $\text{bin}(n_i, \pi_i)$ at x_i , and let $p_i = y_i/n_i$. For binomial GLMs with logit link:

a. For p_i near π_i , show that

$$\log \frac{p_i}{1 - p_i} \approx \log \frac{\pi_i}{1 - \pi_i} + \frac{p_i - \pi_i}{\pi_i(1 - \pi_i)}.$$

b. Show that $z_i^{(t)}$ in (5.24) is a linearized version of the i th sample logit, evaluated at approximation $\pi_i^{(t)}$ for $\hat{\pi}_i$.

c. Verify the formula (5.21) for $\widehat{\text{cov}}(\hat{\beta})$.

¹From J. Henretta, *Social Forces* **66**: 520–536, 1987.

CHAPTER 6

Building, Checking, and Applying Logistic Regression Models

Having studied the basics of fitting and interpreting logistic regression models, we now turn our attention to building and applying them. With several explanatory variables, there are many potential models. In Section 6.1 we discuss strategies for model selection. After choosing a preliminary model, model checking addresses whether systematic lack of fit exists. Section 6.2 covers diagnostics, such as residuals, for model checking. Section 6.3 presents ways of summarizing the predictive power of a model.

In practice, an important application is comparing two groups on a binary response, while adjusting for possibly confounding variables. In Section 6.4 we present the Cochran–Mantel–Haenszel test, a popular way to do this by forming strata for levels of control variables. We then present ways of summarizing the effect, with application to meta-analyses.

Infinite estimates of logistic regression model parameters can occur with certain data configurations. Section 6.5 discusses ways to detect and deal with them. Section 6.6 covers power and sample size determination for logistic regression.

6.1 STRATEGIES IN MODEL SELECTION

Model selection for logistic regression faces the same issues as for ordinary regression. The selection process becomes harder as the number of explanatory variables increases, because of the rapid increase in possible effects and interactions. There are two competing goals: The model should be complex enough to fit the data well. On the other hand, ideally it should be relatively simple to interpret, smoothing rather than overfitting the data. Complications can arise because of the binary nature of the response variable, such as infinite ML parameter estimates for some models when one response outcome is much more common than the other.

Most research studies are designed to answer certain questions. Those questions guide the choice of model terms. Confirmatory analyses then use a restricted set of models. For instance, a study hypothesis about an effect may be tested by comparing models with and without that effect. For studies that are exploratory rather than confirmatory, a search among possible models may provide clues about the dependence structure and raise questions for future research. In either case, it is helpful first to study the effect of each predictor by itself using graphics (incorporating smoothing) for a continuous predictor or conditional distributions within a contingency table for a discrete predictor. This gives you a feel for the marginal effects.

6.1.1 How Many Explanatory Variables Can Be in the Model?

Unbalanced data, with relatively few responses of one type, limit the number of predictors for which we can effectively estimate effects. One guideline based on a Monte Carlo study (Peduzzi et al. 1996) suggested that when there are fewer than 10 outcomes of each type per predictor, impacts can include severely biased parameter estimates, poor standard error estimates, and error rates for Wald tests and confidence intervals far from the nominal level. If $y = 1$ only 30 times out of $n = 1000$, for instance, this guideline implies that ideally the model should contain no more than three predictors.

This is merely one guideline and does *not* mean that you should never consider models that violate it. Many data sets now have large numbers of variables relative to the sample size. With certain strategies presented in Chapter 7, such as penalized likelihood methods that can shrink many estimates to 0, it is possible to have very many predictors. Likewise, you should not use such a guideline to justify being overly ambitious. For example, if you have 1000 outcomes of each type, you are not usually well served by a model with 100 predictors.

Many model selection procedures exist, no one of which is always best. Cautions that apply to ordinary regression hold for any generalized linear model. For instance, a model with several explanatory variables may exhibit *multicollinearity*—correlations among them making it seem that no one variable is important when all the others are in the model. A variable may seem to have little effect because it overlaps considerably with the other explanatory variables in the model, itself being predicted well by the others. Deleting such a redundant variable can be helpful, for instance, to reduce standard errors of other estimated effects.

6.1.2 Example: Horseshoe Crab Mating Data Revisited

The horseshoe crab data set in [Table 4.3](#) has four explanatory variables: color (four categories), spine condition (three categories), weight, and width of the shell. We now fit a logistic regression model using all these to predict whether the female crab has male satellites nearby ($y = 1$).

We start by fitting a model containing all the main effects,

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 \text{weight} + \beta_2 \text{width} + \beta_3 c_1 \\ + \beta_4 c_2 + \beta_5 c_3 + \beta_6 s_1 + \beta_7 s_2,$$

treating color (c_i) and spine condition (s_j) as qualitative (factors), with indicator variables for the first three colors and the first two spine conditions. [Table 6.1](#) shows results. A likelihood-ratio test that Y is jointly independent of these predictors simultaneously tests $H_0: \beta_1 = \dots = \beta_7 = 0$. The test statistic equals 40.56 with $\text{df} = 7$ ($P < 0.0001$). This shows extremely strong evidence that at least one predictor has an effect.

Table 6.1 Software Output (Based on SAS) from Fitting Model with All Main Effects to Horseshoe Crab Data

Testing Global Null Hypothesis: BETA = 0				
Test	Chi-Square	DF	Pr > ChiSq	
Likelihood Ratio	40.5565	7	< .0001	
Analysis of Maximum Likelihood Estimates				
Parameter	Estimate	Std Error	Chi-Square	Pr > ChiSq
Intercept	-9.2734	3.8378	5.8386	0.0157
weight	0.8258	0.7038	1.3765	0.2407
width	0.2631	0.1953	1.8152	0.1779
color 1	1.6087	0.9355	2.9567	0.0855
color 2	1.5058	0.5667	7.0607	0.0079
color 3	1.1198	0.5933	3.5624	0.0591
spine 1	-0.4003	0.5027	0.6340	0.4259
spine 2	-0.4963	0.6292	0.6222	0.4302

Although the overall test is highly significant, the [Table 6.1](#) results are discouraging. The estimates for weight and width are only slightly larger than their SE values. The estimates for the factors compare each category to the final one as a baseline. For color, only one effect is clearly significant; for spine condition, the largest difference is less than a standard error.

The small P -value for the overall test, yet the lack of significance for individual effects, is a warning sign of multicollinearity. In Section 5.2.2 we showed strong evidence of a width effect. Adjusting for weight, color, and spine condition, little evidence remains of a partial width effect. However, weight and width have a strong correlation (0.887). For practical purposes they are equally good predictors, but it is nearly redundant to use them both. Our further analysis uses width (W) with color (C) and spine condition (S) as explanatory variables. For simplicity, we symbolize models by their highest-order terms, regarding C and S as factors. For instance, $(C + S + W)$ denotes a model with main effects, whereas $(C + S * W)$ denotes a model that has those main effects plus an $S \times W$ interaction. It is not usually sensible to consider a model with interaction that does not also contain the main effects that make up that interaction.

6.1.3 Stepwise Procedures: Forward Selection and Backward Elimination

In exploratory studies, an algorithmic method for searching among models can be informative if we use results cautiously. Goodman (1971a) proposed methods analogous to forward selection and backward elimination in ordinary regression.

Forward selection adds terms sequentially. At each stage it selects the term giving the greatest improvement in fit. The minimum P -value for testing the term in the model is a sensible criterion, since reductions in deviance for different terms may have different df values. A point of diminishing returns occurs in adding predictors, when new predictors are so correlated with ones already used that they do not improve predictive power. The process stops when further additions do not significantly improve the fit. A stepwise variation of this procedure retests, at each stage, terms added at previous stages to see if they are still significant.

Backward elimination begins with a complex model and sequentially removes terms. At each stage, it selects the term whose removal has the least damaging effect on the model (e.g., largest P -value). The process stops when any further deletion leads to a significantly poorer fit. With either approach, for qualitative predictors with more than two categories, the process should consider the entire variable at any stage rather than just individual indicator variables. Add or drop the entire variable rather than just one of its indicators. Otherwise, the result depends on the choice of baseline for the indicator coding. The same remark applies to interactions containing that variable.

Some statisticians prefer backward elimination over forward selection, feeling it safer to delete terms from an overly complex model than to add terms to an overly simple one. Forward selection can stop prematurely because a particular test in the sequence has low power. Neither strategy necessarily yields a meaningful model. Use variable selection procedures with caution! Various studies have shown their limitations and pitfalls (e.g., Steyerberg et al. 2001). When you evaluate many terms, one or two that are not truly important may look impressive merely due to chance. For instance, when all the true effects are weak, the largest sample effect is likely to overestimate substantially its true effect. It is best to use such algorithms in an informal manner. This includes the interpretation of P -values used as cutoff points, since the distribution of the minimum or maximum P -value evaluated over a set of predictors is not the same as that of a P -value for a preselected variable.

Some software has additional options for selecting a model. One approach attempts to determine the best model with some fixed number of terms, according to some criterion. If such a method and backward and forward selection procedures yield quite different models, this is an indication that such results are of dubious use. Another such indication would be when a quite different model results from applying a given procedure to a bootstrap sample of the same size from the sample distribution.

Finally, statistical significance should not be the sole criterion for inclusion of a term in a model, and true significance can be difficult to judge in any case (Westfall and Young 1993). It is sensible to include a variable that is central to the purposes of the study and report its estimated effect even if it is not statistically significant. Keeping it in the model may help reduce bias in estimated effects of other predictors and may make it possible to compare results with other studies where the effect is significant, perhaps because of a larger sample size. Algorithmic selection procedures are no substitute for careful thought in guiding the formulation of models.

6.1.4 Example: Backward Elimination for Horseshoe Crab Data

[Table 6.2](#) summarizes results of fitting and comparing several logistic models to the horseshoe crab data with predictors width, color, and spine condition. The deviance (G^2) test of fit compares the model to the saturated model. As noted in Sections 5.2.4 and 5.2.5, this is not approximately chi-squared when a predictor is continuous, as width is. However, the deviance difference between two models that differ by a modest number of parameters is relevant. That difference is the likelihood-ratio statistic $-2(L_0 - L_1)$ comparing the models, and it has an approximate null chi-squared distribution.

Table 6.2 Results of Fitting Several Logistic Regression Models to Horseshoe Crab Data

Model	Predictors ^a	Deviance G^2	df	AIC	Models Compared	Deviance Difference	Corr. $R(y, \hat{\mu})$
1	($C * S * W$)	170.44	152	212.4	—	—	
2	($C * S + C * W + S * W$)	173.68	155	209.7	(2)–(1)	3.2 (df = 3)	
3a	($C * S + S * W$)	177.34	158	207.3	(3a)–(2)	3.7 (df = 3)	
3b	($C * W + S * W$)	181.56	161	205.6	(3b)–(2)	7.9 (df = 6)	
3c	($C * S + C * W$)	173.69	157	205.7	(3c)–(2)	0.0 (df = 2)	
4a	($S + C * W$)	181.64	163	201.6	(4a)–(3c)	8.0 (df = 6)	
4b	($W + C * S$)	177.61	160	203.6	(4b)–(3c)	3.9 (df = 3)	
5	($C + S + W$)	186.61	166	200.6	(5)–(4b)	9.0 (df = 6)	0.456
6a	($C + S$)	208.83	167	220.8	(6a)–(5)	22.2 (df = 1)	0.314
6b	($S + W$)	194.42	169	202.4	(6b)–(5)	7.8 (df = 3)	0.402
6c	($C + W$)	187.46	168	197.5	(6c)–(5)	0.8 (df = 2)	0.452
7a	(C)	212.06	169	220.1	(7a)–(6c)	24.5 (df = 1)	0.285
7b	(W)	194.45	171	198.5	(7b)–(6c)	7.0 (df = 3)	0.402
8	($C = \text{dark} + W$)	187.96	170	194.0	(8)–(6c)	0.5 (df = 2)	0.447
9	None	225.76	172	227.8	(9)–(8)	37.8 (df = 2)	0.000

^a C , color; S , spine condition; W , width.

To select a model, we use backward elimination, at each stage testing only the highest-order terms for each variable. It is inappropriate, for instance, to remove a main effect term if the model has interactions involving that term.

We begin with the most complex model, symbolized by $(C * S * W)$, model 1 in [Table 6.2](#). This model uses main effects for each term as well as the three two-factor interactions and the three-factor interaction. It allows a separate width effect at each CS combination. (In fact, at some of those combinations y outcomes of only one type occur, which implies that those effects are not estimable.) The likelihood-ratio statistic comparing this model to the simpler model $(C * S + C * W + S * W)$ removing the three-factor interaction term equals 3.2 (df = 3). This suggests that the three-factor term is not needed ($P = 0.36$), thank goodness, so we continue the simplification process.

At the next stage we compare the model $(C * S + C * W + S * W)$ to the simpler model $C + S + W$ containing only main effects. The likelihood-ratio statistic comparing the model is the change in deviance, $186.61 - 173.68 = 12.9$ (df = $166 - 155 = 11$). This suggests that two-factor interactions terms are not needed either ($P = 0.30$). [Table 6.2](#) also shows results for intermediate models, and a backward process dropping a term at a time also results in eliminating all the three-factor terms.

At the next stage we consider dropping a main effect term. [Table 6.2](#) shows little consequence of removing S . Both remaining variables (C and W) then have nonnegligible effects. For instance, removing C increases the deviance (comparing models 7b and 6c) by 7.0 on df = 3 ($P = 0.07$). The analysis in Section 5.4.6 revealed a noticeable difference between dark crabs (category 4) and the others. The simpler model that has a single indicator variable for color, equaling 0 for dark crabs and 1 otherwise, fits essentially as well. Further simplification results in large increases in deviance and is unjustified.

6.1.5 Model Selection and the “Correct” Model

In selecting a model from a set of candidates, we are mistaken if we think that there is a “correct” one. Any model is a simplification of reality. For instance, width does not have exactly a linear effect on the probability of satellites, whether we use the logit link or the identity link.

What is the logic of testing the fit of a model when we know that it does not truly hold? A simple model that fits adequately has the advantages of model parsimony. If a model has relatively little bias, describing reality well, it tends to provide more accurate estimates of the quantities of interest.¹

Other criteria besides significance tests can help select a good model in terms of estimating quantities of interest. We next introduce the best known of such criteria.

6.1.6 AIC: Minimizing Distance of the Fit from the Truth

The *Akaike information criterion* (AIC) judges a model by how close its fitted values tend to be to the true mean values, in terms of a certain expected value. Even though a simple model is farther from the true relationship than is a more complex model, it may be preferred because it tends to provide better estimates of certain characteristics, such as cell probabilities. Thus, the optimal model is the one that tends to have fit closest to the true values.

Akaike defined closeness in terms of a Kullback–Leibler measure of distance. Let $p(\mathbf{y})$ denote the probability (or density) of the data under the true model and $p_M(\mathbf{y})$ the probability under the chosen model. The distance measure is $E\{\log[p(\mathbf{y})/p_M(\mathbf{y})]\}$, where the expected value is taken relative to the true distribution. For categorical data, this measure resembles G^2 in form. With a sample, this criterion selects the model that minimizes

$$\text{AIC} = -2(\text{maximized log likelihood} - \text{number of parameters in model}).$$

This penalizes a model for having many parameters. With models for categorical Y , this ordering is equivalent to one based on an adjustment of the deviance, $[G^2 - 2(\text{df})]$, by twice its residual df.

With many potential predictors, we can use the AIC to aid in variable selection. Out of a set of candidate models, we identify the one with smallest AIC. However, models with similar AIC values are also of interest. For instance, we would consider also more parsimonious models that have AIC relatively close to the minimum value.

We illustrate AIC for model selection using the models that [Table 6.2](#) lists. That table also shows the AIC values. Of models using the three basic variables, AIC is smallest (AIC = 197.5) for $C + W$, having main effects of color and width. The simpler model having an indicator variable for whether a crab is dark fares better yet (AIC = 194.0). Either model seems reasonable. We should balance the lower AIC for the simpler model against its having been suggested by the fit of model $C + W$.

An alternative *Bayesian information criterion* (BIC) penalizes more severely for the number of parameters in the model. It replaces 2 by $\log(n)$ as the multiple of the number of parameters, so the selected model is no more complex than the one selected with AIC. Compared with AIC, BIC gravitates less quickly toward more complex models as n increases. It is derived based on a Bayesian argument for determining which of a set of models has highest posterior probability. Differences between BIC values for two models relate to a Bayes factor comparing them. It has the property of selecting the “correct model” with probability converging to 1 as $n \rightarrow \infty$. However, this is based on the Bayesian structure that provides justification for this approach, and its relevance is unclear when applied with frequentist methods. Also, in practice we do not regard any one model as “correct,” so the AIC approach of choosing the model that is closest to reality seems sensible.

For the horseshoe crab mating data, from [Table 6.2](#), AIC = 197.5 for model $(C + W)$ and AIC = 198.5 for model (W) . By contrast, BIC = 213.2 for model $(C + W)$ and BIC = 204.8 for model (W) , thus differing from AIC by preferring the simpler model.

6.1.7 Example: Using Causal Hypotheses to Guide Model Building

Although selection procedures are helpful exploratory tools, the model-building process should utilize theory and common sense. Often, a time ordering among the variables suggests possible causal relationships. Analyzing a certain sequence of models helps to investigate those relationships (Goodman 1973).

We illustrate with [Table 6.3](#), from a British study that employed a random sample survey. A sample of men and women who had petitioned for divorce and an independent sample of married people were asked: (a) “Before you married your (former) husband/wife, had you ever made love with anyone else?”; (b) “During your (former) marriage, (did you have) have you had any affairs or brief sexual encounters with another man/woman?” The $2 \times 2 \times 2 \times 2$ table has variables G = gender, E = extramarital sex report (yes or no), P = premarital sex report, and M = marital status.

Table 6.3 Marital Status by Report of Pre- and Extramarital Sex (PMS and EMS)

		Gender							
		Women				Men			
		Yes		No		Yes		No	
Marital Status	PMS:	Yes	No	Yes	No	Yes	No	Yes	No
	EMS:								
Divorced		17	54	36	214	28	60	17	68
Still married		4	25	4	322	11	42	4	130

Source: G. N. Gilbert, *Modelling Society*. London: George Allen & Unwin, 1981. Reprinted with permission from Unwin Hyman Ltd.

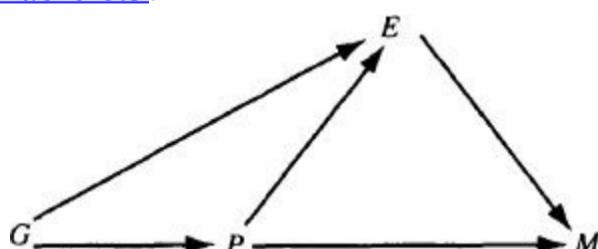
The time points at which responses on the four variables occur suggests the following ordering of the variables:

$$G \rightarrow P \rightarrow E \rightarrow M$$

gender premarital sex extramarital sex marital status

Any of these is an explanatory variable when a variable listed to its right is the response. [Figure 6.1](#) shows one possible causal structure. In this figure, a variable at the tip of an arrow is a response for a model at some stage. The explanatory variables have arrows pointing toward the response, directly or indirectly.

Figure 6.1 Causal diagram for [Table 6.3](#).



We first treat P as a response. [Figure 6.1](#) predicts that G has a direct effect on P , so the model of independence of these variables is inadequate. At the second stage, E is the response. [Figure 6.1](#) predicts that P and G have direct effects on E . It also suggests that G has an indirect effect on E , through its effect on P . These effects on E can be analyzed using the logistic model for E with additive G and P effects. If G has only an indirect effect on E , the model with P alone as a predictor is adequate; that is, at a given level of P , E and G are conditionally independent. At the third stage, M is the response. [Figure 6.1](#) predicts that E has a direct effect on M , P has direct effects and indirect effects through its effects on E , and G has indirect effects through its effects on P and E . This suggests the logistic model for M having additive E and P effects. For this model, G and M are independent, given P and E .

[Table 6.4](#) shows results. The first stage, having P as the response, shows strong evidence of a GP association. The sample odds ratio for their marginal table is 0.27; the estimated odds of premarital sex for females are 0.27 times that for males. The second stage has E as the response. Only weak evidence occurs that G had a direct as well as an indirect effect on E , as G^2 drops by 2.9 ($df = 1$)

after adding G to a model already containing P as a predictor. For this model, the estimated EP conditional odds ratio is 3.6.

Table 6.4 Goodness of Fit of Various Models for Table 6.3^a

Stage	Response Variable	Potential Explanatory	Actual Explanatory	G^2	df
1	P	G	None	75.3	1
			(G)	0.0	0
2	E	G, P	None	48.9	3
			(P)	2.9	2
			($G + P$)	0.0	1
3	M	G, P, E	($E + P$)	18.2	5
			($E * P$)	5.2	4
			($E * P + G$)	0.7	3

^a P , premarital sex; E , extramarital sex; M , marital status; G , gender.

The third stage has M as the response. [Figure 6.1](#) specifies the logistic model with main effects of E and P , but it fits poorly. The model that allows an $E \times P$ interaction in their effects on M but assumes conditional independence of G and M fits much better (G^2 decrease of 13.0, df = 1). The model that also has a main effect for G fits slightly better yet. Either model is more complicated than [Figure 6.1](#) predicted, since the effects of E on M vary according to the level of P . However, some preliminary thought about causal relationships suggested a model similar to one giving a good fit. We leave it to the reader to estimate and interpret effects for the third stage.

6.1.8 Alternative Strategies, Including Model Averaging

In practice, many models can be consistent with the data. If, as stated in Section 6.1.5, no one of them is “correct,” it is logically inconsistent to choose one model based on its fitting the data well and then make subsequent inferences acting as if the model is fixed. This can result in a tendency to underestimate uncertainty and to exaggerate significance. Copas and Eguchi (2010) discussed this issue. They noted that an increasingly popular way of dealing with this is Bayesian model averaging: Identify a set of plausible models, specify prior probabilities for them, and base inference on a weighting according to posterior model probabilities. Copas and Eguchi proposed an alternative approach that identifies statistically equivalent models (that are consistent with the data) and constructs an “envelope likelihood” that reflects the model uncertainty. For estimation of a particular measure, this approach typically generates wider limits that more appropriately reflect the uncertainty.

As computing power continues to explode, enormous data sets are more common, in applications as diverse as genomic investigations and credit scoring by financial institutions. Many applications have huge numbers of potential explanatory variables, making model selection much more difficult. We discuss special issues for such cases in Section 7.5.

In summary, although the focus of this section has been “model selection,” it is often not sensible to have the goal of picking a single model. Also, we should keep in mind the selection uncertainty when we make inferences based on a model, and also realize the tentative nature of using the same data in making those inferences that were used to select a model.

6.2 LOGISTIC REGRESSION DIAGNOSTICS

In Section 5.2.3 we introduced statistics for checking model fit in a global sense. After selecting a preliminary model, we obtain further insight by switching to a microscopic mode of analysis. In contingency tables, for instance, the pattern of lack of fit revealed in cell-by-cell comparisons of observed and fitted counts may suggest a better model or may indicate a segment of the population for which a generally good-fitting model fails.

6.2.1 Residuals: Pearson, Deviance, and Standardized

With categorical predictors, it is useful to form residuals to compare observed and fitted counts. Let y_i denote the binomial outcome for n_i trials at setting i of the explanatory variables, $i = 1, \dots, N$. let $\hat{\pi}_i$ denote the model estimate of $P(Y=1)$. Then $\hat{\mu}_i = n_i \hat{\pi}_i$ is the fitted number of successes.

For a GLM with binomial random component, for observation i the Pearson residual (4.41) is

$$(6.1) \quad e_i = \frac{y_i - n_i \hat{\pi}_i}{\sqrt{\text{var}(Y_i)}} = \frac{y_i - n_i \hat{\pi}_i}{\sqrt{[n_i \hat{\pi}_i (1 - \hat{\pi}_i)]}}.$$

This divides the raw residual $(y_i - \hat{\mu}_i)$ by the estimated binomial standard deviation of y_i . The Pearson statistic for testing the model fit satisfies

$$X^2 = \sum_{i=1}^N e_i^2.$$

An alternative residual uses components of the G^2 fit statistic. This is the *deviance residual*, introduced for GLMs in (4.42). For a binomial GLM, this is

$$(6.2) \quad \sqrt{d_i} \times \text{sign}(y_i - n_i \hat{\pi}_i),$$

where

$$d_i = 2 \left(y_i \log \frac{y_i}{n_i \hat{\pi}_i} + (n_i - y_i) \log \frac{n_i - y_i}{n_i - n_i \hat{\pi}_i} \right).$$

As explained in Section 4.5.6, these and the $\{e_i\}$ are less variable than $N(0, 1)$.

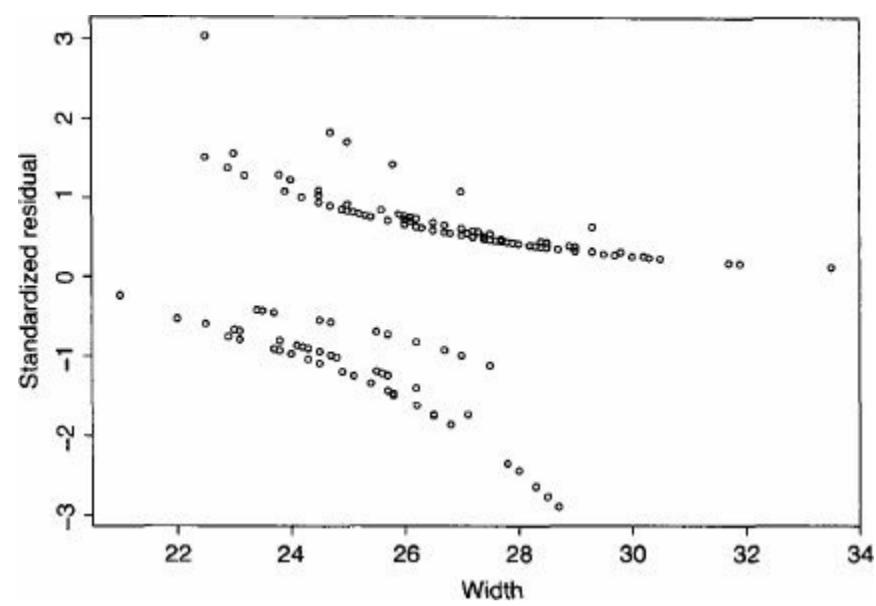
A standardized version of the Pearson residual divides it by its estimated standard error. As noted in Section 4.5.6, this is larger than the Pearson residual, with adjustment that uses the leverage from an estimated hat matrix. For observation i with leverage \hat{h}_i , the standardized residual is

$$r_i = \frac{e_i}{\sqrt{1 - \hat{h}_i}} = \frac{y_i - n_i \hat{\pi}_i}{\sqrt{[n_i \hat{\pi}_i (1 - \hat{\pi}_i)(1 - \hat{h}_i)]}}.$$

It has the advantages compared with the Pearson and deviance residuals of having an approximate $N(0, 1)$ distribution when the model holds and appropriately recognizing redundancies (as noted for 2×2 tables in Section 3.3.1 and in Section 6.2.3 below). Absolute values larger than roughly 2 or 3 provide evidence of lack of fit. It takes larger values to be noteworthy when relatively more of them are inspected.

Plots of residuals against explanatory variables or linear predictor values may detect a type of lack of fit. When fitted values are very small, however, just as X^2 and G^2 lose relevance, so do residuals. When explanatory variables are continuous, often $n_i = 1$ at each setting. Then y_i can equal only 0 or 1, and e_i can assume only two values. One must then be cautious about regarding either outcome as extreme, and a single residual is usually uninformative (see Exercise 6.32). Plots of residuals also then have limited use. Figure 6.2 illustrates, plotting for the horseshoe crab data the standardized residuals against width for the model (5.13) fitted in Section 5.4.5 having width and color as predictors. Width has a strong positive effect, so necessarily for small width values an observation of $y = 1$ will have a relatively large positive residual whereas for large width values an observation of $y = 0$ will have a relatively large negative residual. When plotted against fitted values, a plot of the raw residuals consists merely of two parallel lines of points. The deviance itself is then completely uninformative (Exercise 5.35). When data can be grouped into sets of observations having common predictor values, it is better to compute residuals for the grouped data than for individual subjects.

Figure 6.2 Plot of standardized residuals against width, for model predicting horseshoe crab satellites using width and color predictors.



6.2.2 Example: Heart Disease and Blood Pressure

A sample of male residents of Framingham, Massachusetts, aged 40 through 59, were classified on several factors, including systolic blood pressure. The response variable is whether they developed coronary heart disease during a six-year follow-up period. [Table 6.5](#) shows results.

Table 6.5 Presence of Heart Disease by Blood Pressure, with Fit of Logistic Models and Standardized Residuals

Systolic Pressure (mmHg)	Sample Size	Observed Heart Disease	Fitted		Standardized Residual	
			Independence Model	Linear Logit	Independence Model	Linear Logit
<117	156	3	10.8	5.2	-2.62	-1.11
117–126	252	17	17.4	10.6	-0.12	2.37
127–136	284	12	19.7	15.1	-2.02	-0.95
137–146	271	16	18.8	18.1	-0.74	-0.57
147–156	139	12	9.6	11.6	0.84	0.13
157–166	85	8	5.9	8.9	0.93	-0.33
167–186	99	16	6.9	14.2	3.76	0.65
>186	43	8	3.0	8.4	3.07	-0.18

Source: Data from Cornfield (1962).

Let π_i be the probability of heart disease for blood pressure category i . The table shows the fit and the standardized residuals for two logistic regression models. The first model,

$$\text{logit}(\pi_i) = \alpha,$$

treats the response as independent of blood pressure. Some residuals for that model are large. This is not surprising, since the model fits poorly ($G^2 = 30.02$, $X^2 = 33.38$, $df = 7$).

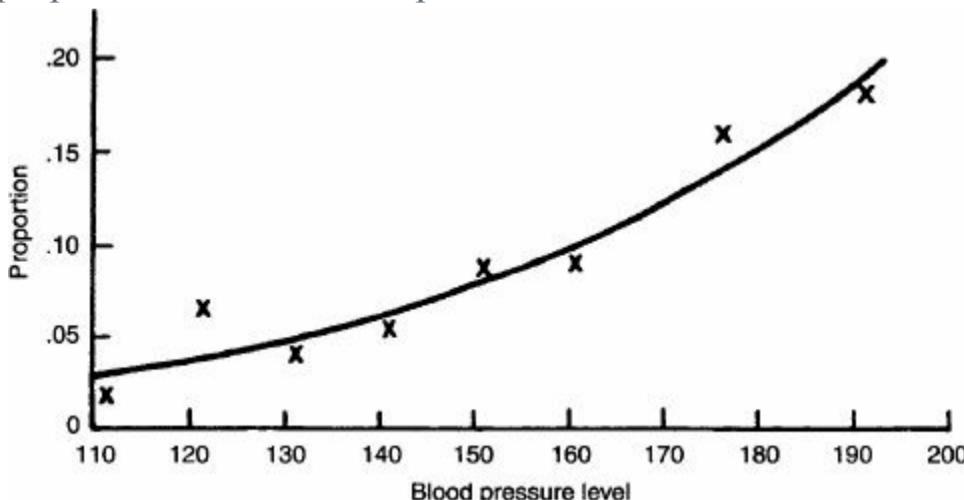
A plot of the residuals for the independence model shows an increasing trend. This suggests the linear logit model,

$$\text{logit}(\pi_i) = \alpha + \beta x_i,$$

with scores $\{x_i\}$ for systolic blood pressure level. We used scores (111.5, 121.5, 131.5, 141.5, 151.5, 161.5, 176.5, 191.5). The nonextreme scores are midpoints for the intervals of blood pressure. The trend in standardized residuals disappears for this model, and only the second category shows some evidence of lack of fit. A single relatively large residual is not surprising, however. With many residuals, a few may be large merely by chance. Here the overall fit statistics ($G^2 = 5.91$, $X^2 = 6.29$, with $df = 6$) do not indicate problems. In analyzing residual patterns, we should be cautious about attributing patterns to what might be chance variation from a model.

A useful graphical display for showing lack of fit compares sample and fitted proportions by plotting them against each other or by plotting both of them against explanatory variables. For the linear logit model, [Figure 6.3](#) plots both the sample proportions and the estimated probabilities of heart disease against blood pressure. The fit seems decent.

Figure 6.3 Sample proportions and estimated probabilities of heart disease for linear logit model.



Studying residuals helps us understand either why a model fits poorly or where there is lack of fit in a generally good-fitting model. The next example illustrates the second case.

6.2.3 Example: Admissions to Graduate School at Florida

[Table 6.6](#) refers to graduate school applications for the 23 departments in the College of Liberal Arts and Sciences at the University of Florida during the 1997–1998 academic year. It cross-classifies the applicant's gender, department to which he or she applied, and whether he or she was admitted, which we treat as the response variable. For gender i in department k , let y_{ik} denote the number admitted and let π_{ik} denote the probability of admission. We treat $\{Y_{ik}\}$ as independent $\text{bin}(n_{ik}, \pi_{ik})$. Other things being equal, we would hope the admissions decision is independent of gender. The model with no gender effect, given the department, is

Table 6.6 Graduate School Admissions by Gender and Department, with Standardized Residuals for Model of No Gender Effect

Dept	Females		Males		Std. Res (Fem, Yes)	Dept	Females		Males		Std. Res (Fem, Yes)
	Yes	No	Yes	No			Yes	No	Yes	No	
anth	32	81	21	41	-0.76	ling	21	10	7	8	1.37
astr	6	0	3	8	2.87	math	25	18	31	37	1.29
chem	12	43	34	110	-0.27	phil	3	0	9	6	1.34
clas	3	1	4	0	-1.07	phys	10	11	25	53	1.32
comm	52	149	5	10	-0.63	poli	25	34	39	49	-0.23
comp	8	7	6	12	1.16	psyc	2	123	4	41	-2.27
engl	35	100	30	112	0.94	reli	3	3	0	2	1.26
geog	9	1	11	11	2.17	roma	29	13	6	3	0.14
geol	6	3	15	6	-0.26	soci	16	33	7	17	0.30
germ	17	0	4	1	1.89	stat	23	9	36	14	-0.01
hist	9	9	21	19	-0.18	zool	4	62	10	54	-1.76
lati	26	7	25	16	1.65						

Source: Data courtesy of Prof. James Booth.

$$\text{logit}(\pi_{ik}) = \alpha + \beta_k^D.$$

However, this model fits rather poorly ($G^2 = 44.74$, $X^2 = 40.85$, $\text{df} = 23$).

The software output in [Table 6.6](#) reports standardized residuals $\{r_i\}$ for the number of females who were admitted. For instance, the Astronomy department admitted 6 females, which was 2.87 estimated standard deviations higher than the model predicted. Each department has only a single nonredundant standardized residual, because of marginal constraints for the model. The model has fit $\hat{\pi}_{ik} = (y_{1k} + y_{2k})/n_{+k}$, corresponding to an independence fit ($\hat{\pi}_{1k} = \hat{\pi}_{2k}$) in each partial table. Now,

$$y_{1k} - n_{1k}\hat{\pi}_{1k} = y_{1k} - n_{1k} \frac{(y_{1k} + y_{2k})}{n_{+k}} = \frac{n_{2k}}{n_{+k}}y_{1k} - \frac{n_{1k}}{n_{+k}}y_{2k} = -(y_{2k} - n_{2k}\hat{\pi}_{2k}).$$

Thus, standard errors of $(y_{1k} - n_{1k}\hat{\pi}_{1k})$ and $(y_{2k} - n_{2k}\hat{\pi}_{2k})$ are identical. The standardized residuals are identical in absolute value for males and females but of different sign. Astronomy admitted 3 males, and their standardized residual was -2.87; the number admitted was 2.87 estimated standard deviations lower than predicted.

Having a single nonredundant value r_i for each df is an advantage of standardized residuals over Pearson (or deviance) residuals. The model of conditional independence has $\text{df} = 1$ for each partial table. Only one bit of information exists about how the data depart from the model, yet the Pearson residual for males need not equal the Pearson residual for females in absolute value. The $\{r_i\}$ for females who were admitted in each department satisfy $\sum_{i=1}^{23} r_i^2 = X^2$, their squares giving 23 df = 1 components for the Pearson statistic. The 46 squared Pearson residuals would have the same sum, but each has null distribution smaller than χ^2_1 .

Departments with large standardized residuals reveal the reason for the lack of fit. Significantly more females were admitted than the model predicts in the Astronomy and Geography departments, and fewer in the Psychology department. Without these three departments, the model fits reasonably well ($G^2 = 24.37$, $X^2 = 22.75$, $\text{df} = 20$).

For the complete data, adding a gender effect to the model does not provide an improved fit ($G^2 = 42.36$, $X^2 = 38.99$, $\text{df} = 22$), because the departments just mentioned have associations in different directions and of greater magnitude than other departments. This model has an ML estimate of 1.19 for the gender conditional odds ratio, the odds of admission being 19% higher for females than

males, given department. By contrast, the marginal table collapsed over department has a sample odds ratio of 0.94, the overall odds of admission being 6% lower for females. This illustrates Simpson's paradox (Section 2.3.2), the estimated conditional association having different direction than the estimated marginal association.

6.2.4 Influence Diagnostics for Logistic Regression

Other regression diagnostic tools are also helpful in assessing fit. These include plots of ordered standardized residuals against normal percentiles (Haberman 1973a) and analyses that describe an observation's influence on parameter estimates and fit statistics. Whenever a residual indicates that a model fits an observation poorly, it can be informative to delete the observation and refit the model to remaining ones. This is equivalent to adding a parameter to the model for that observation, forcing a perfect fit for it.

For ungrouped binary data, the notion of an outlier is not as clear as in ordinary regression. Copas (1988) used a probabilistic definition whereby, if the fitted model were true, the observation would be very unlikely to occur. But then, if $\hat{\pi}_i$ is close to 1 or close to 0 over certain regions of explanatory variable values, it is not at all surprising to observe some outliers. Copas studied how various models differ in their sensitivity to outliers.

As in ordinary regression, a single observation can be quite influential in determining parameter estimates. The greater an observation's leverage, the greater its potential influence. The fit could be quite different if an observation that appears to be an outlier on y and has large leverage is deleted. However, a single observation can have a much more exorbitant influence in ordinary least-squares regression than in logistic regression, since ordinary regression has no bound on the distance of y_i from its expected value. In Section 4.5.6 we observed that the GLM estimated hat matrix

$$\hat{H}_{at} = \hat{W}^{1/2} \mathbf{X} (\mathbf{X}' \hat{W} \mathbf{X})^{-1} \mathbf{X}' \hat{W}^{1/2}$$

depends on the fit as well as the model matrix \mathbf{X} . For logistic regression, recall (from Section 5.5.2) that the weight matrix \hat{W} is diagonal with element $\hat{w}_i = n_i \hat{\pi}_i (1 - \hat{\pi}_i)$ for the n_i observations at setting i of predictors. Points that have extreme predictor values need not have high leverage. In fact, the leverage can be relatively small if $\hat{\pi}_i$ is close to 0 or 1.

Several measures describe the effect of removing an observation from the data set. They are related algebraically to the observation's leverage (Pregibon 1981, Williams 1987). In logistic regression, the observation could be a single binary response or a binomial response for a set of subjects all having the same predictor values (i.e., *ungrouped* or *grouped* data). For each observation, influence measures of deleting the observation include:

1. For each model parameter, the change in its estimate. This change, divided by its standard error, is called *Dfbeta*.
2. A measure of the change in a joint confidence interval for the parameters. This confidence interval displacement diagnostic is denoted by c .
3. The change in X^2 or G^2 goodness-of-fit statistics. Pregibon (1982) showed that the change in X^2 approximates the squared standardized residual for that observation.

For each measure, the larger the value, the greater the influence. With continuous or multiple predictors, it can be informative to plot these diagnostics, for instance, against the estimated probabilities.

We illustrate the diagnostics using the linear logit model for [Table 6.5](#), which has blood pressure as a predictor for heart disease. [Table 6.7](#) contains simple approximations (due to Pregibon 1981) for the *Dfbeta* measure for the coefficient of blood pressure, the confidence interval diagnostic c , the change in G^2 , and the change in X^2 (which is the square of the standardized residual, r_i^2). All their values show that deleting the second observation has the greatest effect. This is not surprising, as that observation has the only relatively large residual. By contrast, [Table 6.7](#) also contains the changes in X^2 and G^2 for deleting observations in fitting the independence model. At the low and high ends of the blood pressure values, several changes are very large. However, these all relate to removing an entire binomial sample at a blood pressure level instead of removing a single subject's binary observation. Such subject-level (ungrouped data) deletions have little effect even for this model.

Table 6.7 Diagnostic Measures for Logistic Regression Models Fitted to Heart Disease Data

Blood Pressure	<i>Dfbeta</i>	<i>c</i>	Pearson X^2 Diff.	Likelihood-Ratio G^2 Diff.	Pearson X^2 Diff. ^a	Likelihood-Ratio G^2 Diff. ^a
111.5	0.49	0.34	1.22	1.39	6.86	9.13
121.5	-1.14	2.26	5.64	5.04	0.02	0.02
131.5	0.33	0.31	0.89	0.94	4.08	4.56
141.5	0.08	0.09	0.33	0.34	0.55	0.57
151.5	0.01	0.00	0.02	0.02	0.70	0.66
161.5	-0.07	0.02	0.11	0.11	0.87	0.80
176.5	0.40	0.26	0.42	0.42	14.17	10.83
191.5	-0.12	0.02	0.03	0.03	9.41	6.73

^aIndependence model; other values refer to linear logit model with blood pressure predictor.

6.3 SUMMARIZING THE PREDICTIVE POWER OF A MODEL

In ordinary regression, R^2 describes the reduction in the conditional variation of the response compared with the marginal variation. It and the multiple correlation R describe how well the explanatory variables can predict the response, with $R = 1$ for perfect prediction. Despite various attempts to define analogs for categorical response models, no proposed measure is as widely useful as R and R^2 . In this section we present a few ways proposed for summarizing predictive power.

6.3.1 Summarizing Predictive Power: R and R -Squared Measures

For any GLM, the correlation $R(y, \hat{\pi})$ between the observed responses $\{y_i\}$ and the model's fitted values $\{\hat{\pi}_i\}$ measures predictive power. For least-squares regression, R is the multiple correlation between Y and the predictors. An advantage of the correlation, relative to its square, is the appeal of working on the original scale and its approximate proportionality to effect size: For a small effect with a single predictor, doubling the slope corresponds approximately to doubling R .

In logistic regression with ungrouped data, $\hat{\mu}_i$ for a particular model is the estimated probability $\hat{\pi}_i$ for binary observation i . So, $R(y, \hat{\mu})$ is then the correlation between the n binary $\{y_i\}$ observations (1 or 0 for each) and the estimated probabilities. The highly discrete nature of y can suppress the range of possible R values. Nevertheless, R is useful for comparing fits of different models for the same data. A caveat is that with many predictors the R estimates can become highly biased upwards in estimating the true correlation, $R(Y, E(Y|X))$, so it can be misleading to compare sample R values for models with greatly different df values. A jackknife adjustment can reduce this bias (Zheng and Agresti 2000).

Another way to measure the association between the binary responses $\{y_i\}$ and their fitted values $\{\hat{\pi}_i\}$ uses the proportional reduction in squared error

$$1 - \frac{\sum_i (y_i - \hat{\pi}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

obtained by using $\hat{\pi}_i$ instead of $\bar{y} = \sum_j y_j / n$ as a predictor of y_i (Efron 1978). Amemiya (1981) suggested a related measure that weights squared deviations by inverse predicted variances. For logistic regression, unlike normal GLMs, these and $R(y, \hat{\mu})$ need not be nondecreasing as the model gets more complex. Like any correlation-type measure, they can depend strongly on the range of observed values of explanatory variables, and as computed for sample data are biased upward as estimates of corresponding population measures. Bias corrections are possible (e.g., Liao and McGee 2003).

6.3.2 Summarizing Predictive Power: Likelihood and Deviance Measures

Other measures of predictive power directly use the likelihood function. Denote the maximized log likelihood by L_M for a given model, L_S for the saturated model, and L_0 for the null model containing only an intercept term. Probabilities are no greater than 1.0, so log likelihoods are nonpositive. As the model complexity increases, the parameter space expands, so the maximized log likelihood increases. Thus, $L_0 \leq L_M \leq L_S \leq 0$. The measure

$$(6.3) \quad D = \frac{L_M - L_0}{L_S - L_0}$$

falls between 0 and 1. It equals 0 when the model provides no improvement in fit over the null model, and it equals 1 when the model fits as well as the saturated model. A weakness is that the log likelihood is not an easily interpretable scale. Interpreting the numerical value is difficult, other than in a comparative sense for different models.

For N independent Bernoulli observations, the maximized log likelihood is

$$\log \prod_{i=1}^N [\hat{\pi}_i^{y_i} (1 - \hat{\pi}_i)^{1-y_i}] = \sum_{i=1}^N [y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i)].$$

The null model gives $\hat{\pi}_i = (\sum_i y_i)/N = \bar{y}$, so that

$$L_0 = N[\bar{y} \log \bar{y} + (1 - \bar{y}) \log(1 - \bar{y})].$$

The saturated model has a parameter for each subject and implies that $\hat{\pi}_i = y_i$, for all i . Thus, $L_S = 0$ and (6.3) simplifies to

$$D = \frac{L_0 - L_M}{L_0}.$$

McFadden (1974) proposed this measure.

Suppose there are multiple observations at each setting of explanatory variables. Then, the data file can take the grouped-data form of N binomial counts with binomial indices $\{n_i\}$, rather than the ungrouped form of N Bernoulli indicators each with $n_i = 1$. The saturated model then has a parameter for each count. It gives N fitted proportions equal to the N sample proportions of success. Then L_S is nonzero and (6.3) takes a different value than when calculated using individual subjects. For N binomial counts, the maximized likelihoods are related to the G^2 goodness-of-fit statistic by $G^2(M) = -2(L_M - L_S)$, so (6.3) becomes

$$D^* = \frac{G^2(0) - G^2(M)}{G^2(0)}.$$

Goodman (1971a) and Theil (1970) discussed this and related partial association measures.

With grouped data D^* can be large even when predictive power is weak at the subject level. For instance, a model can fit much better than the null model even though fitted probabilities are close to 0.50 for the entire sample. In particular, $D^* = 1$ when it fits perfectly, regardless of how well one can predict individual subjects' responses on Y with that model. Also, suppose that the population satisfies the given model, but not the null model. As the sample size $n = \sum_i n_i$ increases with number of settings N fixed, $G^2(M)$ behaves like a chi-squared random variable but $G^2(0)$ eventually grows unboundedly. Thus, $D^* \rightarrow 1$ (in probability) as $n \rightarrow \infty$, and its magnitude tends to depend on n . This measure confounds model goodness of fit with predictive power. Similar behavior occurs for R^2 in regression analyses when calculated using *means* of y values (rather than individual y values) at N different x settings. It is more sensible to use D for binary, ungrouped data.

6.3.3 Summarizing Predictive Power: Classification Tables

A *classification table* cross-classifies the binary response with a prediction of whether $y = 0$ or 1 . The prediction for observation i is $\hat{y} = 1$ when $\hat{\pi}_i > \pi_0$ and $\hat{y} = 0$ when $\hat{\pi}_i \leq \pi_0$, for some cutoff π_0 . One possibility is $\pi_0 = 0.50$. Another is the sample proportion of 1 outcomes, which is $\hat{\pi}_i$ for the model containing only an intercept term. Rather than using $\hat{\pi}_i$ from the model fitted to the data set of which y_i was one element, it is better to make the prediction with the “leave-one-out” cross-validation approach by which $\hat{\pi}_i$ is based on the model fitted to the other $n - 1$ observations.

Using a classification table, we can summarize the predictive power by

$$\text{sensitivity} = P(\hat{y} = 1 | y = 1) \quad \text{and} \quad \text{specificity} = P(\hat{y} = 0 | y = 0).$$

(Recall Section 2.1.3.) An overall summary of predictor power is the proportion of correct classifications. This estimates

$$\begin{aligned} P(\text{correct classification}) &= P(y = 1 \text{ and } \hat{y} = 1) + P(y = 0 \text{ and } \hat{y} = 0) \\ &= P(\hat{y} = 1 | y = 1)P(y = 1) + P(\hat{y} = 0 | y = 0)P(y = 0), \end{aligned}$$

which is a weighted average of sensitivity and specificity.

A classification table has limitations: It collapses continuous predictive values $\hat{\pi}$ into binary ones. The choice of π_0 is arbitrary. Results are sensitive to the relative numbers of times that $y = 1$ and $y = 0$. For example, if a low proportion of observations have $y = 1$, the model fit may never have $\hat{\pi}_i > 0.50$, in which case one never predicts $\hat{y} = 1$. Again, the main use is for comparing different models with the same data.

6.3.4 Summarizing Predictive Power: ROC Curves

The classification table summaries depend on the cutoff π_0 for making classifications. A *receiver operating characteristic* (ROC) curve is a plot of sensitivity as a function of (1 – specificity) for the possible π_0 . A ROC curve is more informative than a classification table, because it summarizes predictive power for all possible π_0 . When π_0 is near 0, almost all predictions are $\hat{y} = 1$; then, sensitivity is near 1, specificity is near 0, and the point (1 – specificity, sensitivity) $\approx (1, 1)$. When π_0 is near 1, almost all predictions are $\hat{y} = 0$; then, sensitivity is near 0, specificity is near 1, and (1 – specificity, sensitivity) $\approx (0, 0)$. A ROC curve usually has a concave shape connecting the points (0, 0) and (1, 1).

For a given specificity, better predictive power corresponds to higher sensitivity. So, the better the predictive power, the higher the ROC curve. In a summary sense, the greater the area under the ROC curve, the better the predictions. In fact, the area under a ROC curve is identical to the value of another measure of predictive power, the *concordance index* (Hanley and McNeil 1982). Consider all pairs of observations (i, j) for which $y_i = 1$ and $y_j = 0$. The concordance index c is the proportion of such pairs for which $\hat{\pi}_i > \hat{\pi}_j$; that is, it is the relative frequency of the pairwise predictions and the outcomes being concordant, the observation with the larger y also having the larger $\hat{\pi}$. A value $c = 0.50$ means predictions are no better than random guessing. This corresponds to a model having only an intercept term and an ROC curve that is a straight line connecting points (0, 0) and (1, 1).

6.3.5 Example: Evaluating Predictive Power for Horseshoe Crab Data

[Table 6.2](#) shows the correlation $R(y, \hat{y})$ for some models fitted to the horseshoe crab data for predicting whether a female crab had at least one satellite. Color alone (C) has $R = 0.285$, width alone (W) has $R = 0.402$, and using both ($C + W$) increases R to 0.452. The simpler model ($C = \text{dark} + W$) that uses color as binary merely to indicate whether a crab is dark does nearly as well, with $R = 0.447$. These models fit essentially as well as more complex models not shown in the table. For example, the model that adds an interaction term to the model ($C = \text{dark} + W$) has $R = 0.452$.

Other measures of predictive power have different magnitudes but similar results in comparing various models. For example, the concordance index $c = 0.639$ with model (C) (in factor form), 0.742 with model (W), 0.771 with model ($C + W$), 0.772 with model ($C = \text{dark} + W$), and 0.772 for the model that adds an interaction term to this model.

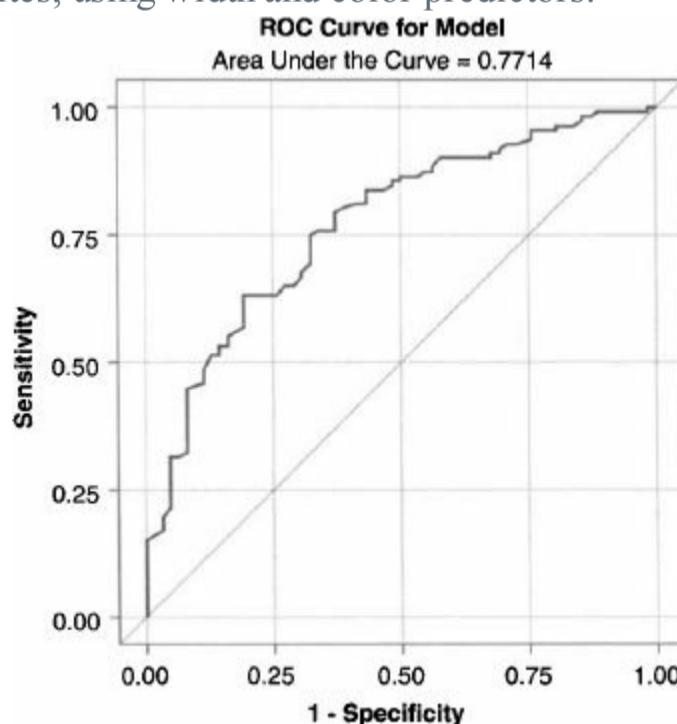
Next, we illustrate a classification table, for the model ($C + W$). Of the 173 crabs, 111 had a satellite, for a sample proportion of 0.642. [Table 6.8](#) shows classification tables using $\pi_0 = 0.50$ and $\pi_0 = 0.642$ with cross-validated predictions. When $\pi_0 = 0.642$, from [Table 6.8](#) the estimated sensitivity = $74/111 = 0.667$ and specificity = $42/62 = 0.677$. The proportion of correct classifications is $(74 + 42)/173 = 0.671$.

Table 6.8 Classification Tables for Horseshoe Crab Mating Data

Actual	Prediction, $\pi_0 = 0.642$		Prediction, $\pi_0 = 0.500$		Total
	$\hat{y} = 1$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 0$	
$y = 1$	74	37	94	17	111
$y = 0$	20	42	34	28	62

[Figure 6.4](#) shows how PROC LOGISTIC in SAS reports the ROC curve for the model ($C + W$). When $\pi_0 = 0.642$, specificity = 0.68, sensitivity = 0.67, and the point plotted has coordinates (0.32,0.67). The area under the curve is $c = 0.771$.

[Figure 6.4](#) ROC curve (from SAS PROC LOGISTIC) for logistic regression model estimating the probability a crab has satellites, using width and color predictors.



6.4 MANTEL-HAENSZEL AND RELATED METHODS FOR MULTIPLE 2×2 TABLES

The analysis of the graduate admissions data in Section 6.2.3 used the model of conditional independence. This model is an important one in biomedical studies that investigate whether an association exists between a treatment variable and a disease outcome after adjusting for a possibly confounding variable that might influence that association. We next present the test of conditional independence as a logistic model analysis for a $2 \times 2 \times K$ contingency table. We also present a test and a related estimation method, due to Mantel and Haenszel (1959), that seem non-model-based but relate to the same logistic model.

We illustrate using [Table 6.9](#), showing results of a clinical trial with eight centers. The study compared two cream preparations, an active drug and a control, on their success in curing an infection. This table illustrates a common pharmaceutical application, comparing two treatments on a binary response with observations from several strata. The strata are often medical centers or clinics; or, they may be levels of age or severity of the condition being treated; or, they may be combinations of levels of several control variables; or, they may be different studies of the same sort summarized in a meta-analysis.

Table 6.9 Clinical Trial Relating Treatment to Response for Eight Centers, with Expected Value and Variance (of Success Count for Drug) Under Conditional Independence

Center	Treatment	Response		Odds Ratio	μ_{11k}	$\text{var}(n_{11k})$
		Success	Failure			
1	Drug	11	25	1.19	10.36	3.79
	Control	10	27			
2	Drug	16	4	1.82	14.62	2.47
	Control	22	10			
3	Drug	14	5	4.80	10.50	2.41
	Control	7	12			
4	Drug	2	14	2.29	1.45	0.70
	Control	1	16			
5	Drug	6	11	∞	3.52	1.20
	Control	0	12			
6	Drug	1	10	∞	0.52	0.25
	Control	0	10			
7	Drug	1	4	2.0	0.71	0.42
	Control	1	8			
8	Drug	4	2	0.33	4.62	0.62
	Control	6	1			

Source: Beitler and Landis (1985).

6.4.1 Using Logistic Models to Test Conditional Independence

For a binary response Y , we analyze the effect of a binary predictor X , conditional on the category of a qualitative covariate Z . Let $\pi_{ik} = P(Y = 1|X = i, Z = k)$. Consider the model

$$(6.4) \text{ logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z, \quad i = 1, 2, \quad k = 1, \dots, K,$$

where $x_1 = 1$ and $x_2 = 0$. This model assumes that the XY conditional odds ratio is the same at each category of Z , namely, $\exp(\beta)$. The null hypothesis of XY conditional independence is $H_0: \beta = 0$. The Wald statistic is $(\hat{\beta}/SE)^2$. The likelihood-ratio statistic is the difference between deviance statistics for the reduced model

$$(6.5) \text{ logit}(\pi_{ik}) = \alpha + \beta_k^Z$$

and the full model. These tests are sensible when X has a similar effect at each category of Z . They have $\text{df} = 1$.

Alternatively, since the reduced model (6.5) is equivalent to conditional independence of X and Y , we can test conditional independence using a goodness-of-fit test of that model. Such a test has $\text{df} = K$ when X is binary. This corresponds to comparing model (6.5) and the saturated model, which permits $\beta \neq 0$ in (6.4) and also contains $(K - 1) XZ$ interaction parameters. The likelihood-ratio test statistic partitions into two components, the likelihood-ratio statistic with $\text{df} = 1$ for testing $H_0: \beta = 0$ in model (6.4) and the likelihood-ratio statistic with $\text{df} = (K - 1)$ for testing the fit of model (6.4) and thus equality of the K odds ratios (Goodman 1969, Cheng et al. 2010).

When no interaction exists or when the conditional XY association has relatively little variation among the levels of Z , it follows from results in Section 5.3.7 that the approach using $\text{df} = K$ of testing conditional independence is less powerful, especially when K is large. When model (6.4) holds, both tests have the same noncentrality. Thus, the test of $\beta = 0$ in model (6.4) is more powerful, since it has fewer degrees of freedom. However, when the direction of the conditional XY association varies among categories of Z , it can be less powerful.

6.4.2 Cochran–Mantel–Haenszel Test of Conditional Independence

Mantel and Haenszel (1959) proposed a non-model-based test of H_0 : conditional independence in $2 \times 2 \times K$ tables. Focusing on retrospective studies of disease, they treated response (column) marginal totals as fixed. Thus, in each partial table k of cell counts $\{n_{ijk}\}$, their analysis conditioned on both the treatment (e.g., group) totals $\{n_{1+k}, n_{2+k}\}$ and the response outcome totals $\{n_{+1k}, n_{+2k}\}$. The usual sampling schemes then yield a hypergeometric distribution (3.17) for the first cell count n_{11k} in each partial table. That count determines $\{n_{12k}, n_{21k}, n_{22k}\}$, given the marginal totals.

Under H_0 , the hypergeometric mean and variance of n_{11k} are

$$\mu_{11k} = E(n_{11k}) = n_{1+k}n_{+1k}/n_{++k},$$

$$\text{var}(n_{11k}) = n_{1+k}n_{2+k}n_{+1k}n_{+2k}/[n_{++k}^2(n_{++k} - 1)].$$

Cell counts from different partial tables are independent. The test statistic combines information from the K tables by comparing $\sum_k n_{11k}$ to its null expected value. It equals

$$(6.6) \quad \text{CMH} = \frac{[\sum_k (n_{11k} - \mu_{11k})]^2}{\sum_k \text{var}(n_{11k})}.$$

This statistic has a large-sample chi-squared null distribution with $\text{df} = 1$.

When the odds ratio $\theta_{XY(k)} > 1$ in partial table k , we expect that $(n_{11k} - \mu_{11k}) > 0$. When $\theta_{XY(k)} < 1$ in every partial table or $\theta_{XY(k)} < 1$ in each table, $\sum_k (n_{11k} - \mu_{11k})$ tends to be relatively large in absolute value. This test works best when the conditional XY association is similar in each partial table. In this sense it is similar to the tests of $H_0: \beta = 0$ in logistic model (6.4). When the sample sizes in the strata are moderately large, this test usually gives similar results. In fact, it is a score test of $H_0: \beta = 0$ in that model (Birch 1964b, 1965, Darroch 1981, Day and Byar 1979).

Cochran (1954) proposed a similar test statistic. He treated the rows in each 2×2 table as two independent binomials rather than a hypergeometric. Cochran's statistic is (6.6) with $\text{var}(n_{11k})$ replaced by

$$\text{var}(n_{11k}) = n_{1+k}n_{2+k}n_{+1k}n_{+2k}/n_{++k}^3.$$

Because of the similarity in their approaches, we call (6.6) the *Cochran–Mantel–Haenszel (CMH) statistic*. The Mantel and Haenszel approach using the hypergeometric is more general in that it also applies to some cases in which the rows are not independent binomial samples from two populations. Examples are (1) retrospective studies and (2) randomized clinical trials with the available subjects (usually volunteers) randomly allocated to two treatments. In the first case the column totals are naturally fixed. In the second, under the null hypothesis the column margins are the same regardless of how subjects are assigned to treatments, and randomization arguments lead to the hypergeometric for each 2×2 table.

Mantel and Haenszel (1959) proposed (6.6) but with a continuity correction. The P -value from the test then better approximates an exact conditional test, based directly on the convolution of the hypergeometric distributions rather than the chi-squared approximation (Section 7.3.5). However, that test tends to be conservative. Mantel and Fleiss (1980) stated that the asymptotic approximation for this test is adequate if the potential values for $\sum_k (n_{11k} - \mu_{11k})$, for the fixed margins in each 2×2 table, can exceed ± 5 . The CMH statistic generalizes for $I \times J \times K$ tables (Section 8.4.3).

6.4.3 Example: Multicenter Clinical Trial Revisited

For the multicenter clinical trial introduced at the beginning of Section 6.4, [Table 6.9](#) reports the sample odds ratio for each table and the expected value and variance of the number of successes for the drug treatment (n_{11k}) under H_0 : conditional independence. In each table except the last, the sample odds ratio shows a positive association. Thus, it makes sense to combine results using CMH = 6.38, with df = 1. There is considerable evidence against H_0 ($P = 0.012$).

Similar results occur in testing $H_0: \beta = 0$ in logistic model [\(6.4\)](#). The model fit has $\hat{\beta} = 0.777$ with $SE = 0.307$. The Wald statistic is $(0.777/0.307)^2 = 6.42$ ($P = 0.011$). The likelihood-ratio statistic equals 6.67 ($P = 0.010$).

6.4.4 CMH Test Is Advantageous for Sparse Data

In summary, for the main-effects logistic model (6.4), the CMH statistic is the score statistic alternative to the likelihood-ratio or Wald test of $H_0: \beta = 0$. As $n \rightarrow \infty$ with fixed K , all three tests have the same asymptotic chi-squared behavior under H_0 . An advantage of the CMH statistic is that its chi-squared limit also applies with an alternative asymptotic scheme in which $K \rightarrow \infty$ as $n \rightarrow \infty$. The asymptotic theory for likelihood-ratio and Wald tests requires the number of parameters (and hence K) to be fixed, so it does not apply to this scheme.

Here is an application of this type: Suppose each stratum has a single matched pair of subjects, one in each group. Then, $n_{1+k} = n_{2+k} = 1$ for each k and $n = 2K$, so $K \rightarrow \infty$ as $n \rightarrow \infty$. [Table 6.10](#) shows the data layout for this situation. When both subjects in stratum k make the same response, as in the first case in [Table 6.10](#), $n_{+1k} = 0$ or $n_{+2k} = 0$. Given the marginal counts, the internal counts are then completely determined, and $\pi_{11k} = n_{11k}$ and $\text{var}(n_{11k}) = 0$. When the subjects make differing responses, as in the second case, $n_{+1k} = n_{+2k} = 1$, so that $\mu_{11k} = 0.50$ and $\text{var}(n_{11k}) = 0.25$. Thus, a matched pair contributes to the CMH statistic only when the two subjects' responses differ. Let K^* denote the number of the K tables that satisfy this. Although each n_{11k} can take only two values, the central limit theorem implies that $\sum_k n_{11k}$ is approximately normal for large K^* . Then, the distribution of CMH is approximately chi-squared.

Table 6.10 Two Examples of a Stratum Containing a Matched Pair

Element of Pair	Response		Response	
	Success	Failure	Success	Failure
First	1	0	1	0
Second	1	0	0	1

Usually, when K grows with n , each stratum has few observations, so the full table is sparse. There may be more than two observations, such as case-control studies that match several controls with each case. The nonstandard setting in which $K \rightarrow \infty$ as $n \rightarrow \infty$ is called *sparse-data asymptotics*. Ordinary ML estimation then breaks down because the number of parameters is not fixed, instead having the same order as the sample size. In particular, the chi-squared approximation is good for the likelihood-ratio and Wald statistics for testing conditional independence when K is fixed and small relative to n and the strata marginal totals mostly exceed about 5 to 10.

6.4.5 Estimation of Common Odds Ratio

It is more informative to estimate the strength of association than to test hypotheses about it. When the association seems stable among partial tables, we can combine the K sample odds ratios into a summary measure of conditional association. The logistic model (6.4) implies homogeneous association, $\theta_{XY(1)} = \dots = \theta_{XY(K)} = \exp(\beta)$. The ML estimate of the common odds ratio is $\exp(\hat{\beta})$.

Other estimators of a common odds ratio are not model-based. Woolf (1955) proposed an exponentiated weighted average of the K sample log odds ratios. Let $p_{ijk} = n_{ijk}/n_{++k}$. Mantel and Haenszel (1959) proposed

$$(6.7) \quad \hat{\theta}_{\text{MH}} = \frac{\sum_k (n_{11k}n_{22k}/n_{++k})}{\sum_k (n_{12k}n_{21k}/n_{++k})} = \frac{\sum_k n_{++k} p_{11|k} p_{22|k}}{\sum_k n_{++k} p_{12|k} p_{21|k}}.$$

This gives more weight to strata with larger sample sizes. With fixed K , $\log(\hat{\theta}_{\text{MH}})$ is slightly less efficient than the ML estimator $\hat{\beta}$ unless $\beta = 0$ (Tarone et al. 1983). However, it is preferred over the ML estimator when K is large and the data are very sparse. The ML estimator $\hat{\beta}$ of the log odds ratio then tends to be too large in absolute value. For sparse-data asymptotics with only a single matched pair in each stratum, for instance, $\hat{\beta} \xrightarrow{P} 2\beta$. (see Exercise 11.29.)

Robins et al. (1986) derived an estimated variance for $\log(\hat{\theta}_{\text{MH}})$ that applies both for standard asymptotics with large n and fixed K and for sparse-data asymptotics in which K is also large. Expressing $\hat{\theta}_{\text{MH}} = R/S = (\sum_k R_k) / (\sum_k S_k)$ with $R_k = n_{11k}n_{22k}/n_{++k}$, their derivation showed that $(\log \hat{\theta}_{\text{MH}} - \log \theta)$ is approximately proportional to $(R - \theta S)$. They also showed that $E(R - \theta S) = 0$ and derived the variance of $(R - \theta S)$. Their result is

$$\begin{aligned} \hat{\sigma}^2[\log \hat{\theta}_{\text{MH}}] &= \frac{1}{2R^2} \sum_k n_{++k}^{-1} (n_{11k} + n_{22k}) R_k \\ &\quad + \frac{1}{2S^2} \sum_k n_{++k}^{-1} (n_{12k} + n_{21k}) S_k \\ &\quad + \frac{1}{2RS} \sum_k n_{++k}^{-1} [(n_{11k} + n_{22k}) S_k + (n_{12k} + n_{21k}) R_k]. \end{aligned}$$

For the eight-center clinical trial summarized by [Table 6.9](#),

$$\hat{\theta}_{\text{MH}} = \frac{\sum_k (n_{11k}n_{22k}/n_{++k})}{\sum_k (n_{12k}n_{21k}/n_{++k})} = \frac{(11 \times 27)/73 + \dots + (4 \times 1)/13}{(25 \times 10)/73 + \dots + (2 \times 6)/13} = 2.13.$$

For $\log \hat{\theta}_{\text{MH}} = 0.758$, $\hat{\sigma}[\log \hat{\theta}_{\text{MH}}] = 0.303$. A 95% confidence interval for the common odds ratio is $\exp(0.758 \pm 1.96 \times 0.303)$ or $(1.18, 3.87)$. Similar results occur using model (6.4). The 95% confidence interval for $\exp(\beta)$ is $\exp(0.777 \pm 1.96 \times 0.307)$, or $(1.19, 3.97)$, using the Wald interval, and $(1.20, 4.02)$ using the likelihood-ratio interval. Although the evidence of an effect is considerable, inference about its size is rather imprecise with such a small sample. The odds of success may be as little as 20% higher with the drug, or they may be as much as four times as high.

If the true odds ratios are not identical but do not vary much, $\hat{\theta}_{\text{MH}}$ still is a useful summary of the conditional associations. Similarly, the CMH test is a powerful summary of evidence against H_0 : conditional independence, as long as the sample associations fall primarily in a single direction. It is not necessary to assume equality of odds ratios to use the CMH test or $\hat{\theta}_{\text{MH}}$.

6.4.6 Meta-analyses for Summarizing Multiple 2×2 Tables

A *meta-analysis* is a statistical analysis that combines information from several studies. For comparing two treatments on a binary response, the analysis refers to a $2 \times 2 \times K$ table, one 2×2 table for each study. For a particular effect measure, such as the odds ratio or a difference of proportions, here we consider the simplifying assumption that the population values of the measure are identical in each study. This is usually unrealistic, but is often adequate for providing a simple summary of the effect when the true effect does not vary much among studies. Sections 6.4.10 and 13.3.6 generalize to allow for heterogeneity among the effects.

Consider first the significance test of the null hypothesis of no effect, that is, conditional independence between the treatment and the response for each study. The logistic model (6.4) is a natural one for such an analysis. We test $H_0: \beta = 0$ using the likelihood-ratio test or the Cochran–Mantel–Haenszel (CMH) test (6.6). As mentioned in Section 6.4.4, the CMH test is advantageous for highly sparse data. When asymptotics are unsuitable even for that test, we can use a small-sample generalization of Fisher’s exact test to multiple 2×2 tables, as presented in Section 7.3.5. For the CMH test or for the small-sample test, tables for which there are either no successes or no failures provide no information about whether there is truly an association and make no contribution to the test. (Recall that Section 6.4.4 discussed this for matched pairs.) There is no reason to use some device such as adding a small constant to cells of the table so those tables enter the analysis, because they are uninformative about the odds ratio (Agresti and Hartzel 2000).

Consider next summarizing the size of the effect. For the logistic model (6.4), we can use the ML estimate of the odds ratio $\exp(\beta)$ and a corresponding confidence interval. For highly sparse data, we can instead use the Mantel–Haenszel estimate $\hat{\beta}_{\text{MH}}$ and its corresponding interval. A small-sample interval can guarantee a lower bound for the coverage probability (Section 16.6.6). For all such frequentist analyses, tables for which there are either no successes or no failures provide no information about the size of the common odds ratio and do not contribute to the estimate.

6.4.7 Meta-analyses for Multiple 2×2 Tables: Difference of Proportions

The difference of proportions and the relative risk are alternative effect measures that are simpler to interpret than the odds ratio. A common difference of proportions for each study is the parameter δ in a model

$$\pi_{ik} = \alpha + \delta x_i + \beta_k^Z, \quad i = 1, 2, \quad k = 1, \dots, K,$$

that replaces the logit link in model (6.4) by the identity link.

Mantel-Haenszel-type estimates are also available for such measures. In stratum k , denote the binomial “success” counts by $s_k = n_{1|k}$ and $t_k = n_{2|k}$ based on sample sizes $m_k = n_{1+k}$ and $n_k = n_{2+k}$, and let $N_k = m_k + n_k$. With $w_k = m_k n_k / N_k$, the estimator of a common difference of proportions is the weighted average of the stratum-specific estimates $\hat{\delta}_k = [(s_k/m_k) - (t_k/n_k)]$,

$$(6.8) \quad \hat{\delta}_{\text{MH}} = \left(\sum_k w_k \hat{\delta}_k \right) / \left(\sum_k w_k \right)$$

(Greenland and Robins 1985). An estimated variance,

$$\hat{\sigma}^2(\hat{\delta}_{\text{MH}}) = \left[\hat{\delta}_{\text{MH}} \left(\sum_k P_k \right) + \left(\sum_k Q_k \right) \right] / \left(\sum_k w_k \right)^2,$$

with

$$P_k = [m_k^2 t_k - n_k^2 s_k + m_k n_k (n_k - m_k)/2] / N_k^2,$$

$$Q_k = [s_k(n_k - t_k) + t_k(m_k - s_k)] / 2N_k,$$

applies under both standard and sparse-data asymptotics (Sato 1989).

Under standard asymptotics, the ML model-based estimator is preferred because it is more efficient. However, ML fitting difficulties often arise when both probabilities are near 0 or near 1, and the $\{\pi_{ik}\}$ must be constrained to fall between 0 and 1. Here is an alternative approach that is then asymptotically efficient and does not have boundary problems: Express the score or profile likelihood $100(1 - \alpha)\%$ confidence interval for the difference of proportions (see Section 3.2.5) for study k alone as $d_k \pm z_{\alpha/2} s_k$, where d_k is the midpoint of that interval (i.e., *not* the sample difference of proportions $\hat{\delta}_k$) and s_k is a “pseudo standard error” that is taken to be the width of the interval divided by $2z_{\alpha/2}$. Then, taking weight $w_k = [1/(s_k^2)] / [\sum_i 1/(s_i^2)]$, we form $\hat{\delta} = \sum_k w_k d_k$, $SE = [\sum_k 1/(s_k^2)]^{-1/2}$, and the summary interval $\hat{\delta} \pm z_{\alpha/2}(SE)$. Unlike Wald methods, this does not require using unreliable sample standard errors from each study but merely uses a midpoint and width based on information obtained from the likelihood function.

To illustrate, the eight-center clinical trial data of Table 6.9 was analyzed in Sections 6.4.3 and 6.4.5 with CMH methods and with logistic model (6.4). For summarizing the effect by a common difference of success proportions between drug and control, the Mantel–Haenszel-type estimate (6.8) is $\hat{\delta}_{\text{MH}} = 0.130$ with $SE = 0.050$. Using the alternative method just mentioned that combines information from the eight center-specific score confidence intervals, we get $\hat{\delta} = 0.128$, $SE = 0.049$, and a 95% confidence interval for a common difference of proportions of $(0.032, 0.224)$.

For the difference of proportions, tables for which there are either no successes or no failures provide no information about whether the true common value δ is nonzero (i.e., the significance testing problem) but they do give information about the magnitude of the effect. If each treatment, for example, has a very large number of failures and no successes, then we have evidence that both population proportions are close to 0 and that the difference is small. Thus, such data do have an impact on practical significance. (See Exercise 6.33 for an illustration.)

Agresti and Hartzel (2000) discussed ways of summarizing information from multiple tables and gave many additional references. Tian et al. (2009) proposed an alternative approach designed for small-sample cases in which some centers may have no outcomes of a particular type.

6.4.8 Collapsibility and Logistic Models for Contingency Tables

We have seen that conditional associations in partial tables usually differ from marginal associations. Under certain *collapsibility conditions* given in Section 2.3.6, however, they are the same. For odds ratios, recall that for three-way tables, XY marginal and conditional odds ratios are identical if either Z and X are conditionally independent or if Z and Y are conditionally independent.

For instance, suppose that a clinical trial studies the association between a binary treatment variable X ($x_1 = 1, x^2 = 0$) and a binary response Y , using data from K centers (Z). The logistic model (6.4), namely,

$$(6.9) \text{ logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z, \quad i = 1, 2, \quad k = 1, \dots, K,$$

has the same treatment effect β for each center. Since the model has no restriction on the conditional association of Z with X or with Y , this effect may differ after collapsing the $2 \times 2 \times K$ table over centers. The estimated XY conditional odds ratio, $\exp(\beta)$, typically differs from the sample odds ratio in the marginal 2×2 table.

Next, consider the simpler model that lacks center effects, $\text{logit}(\pi_{ik}) = \alpha + \beta x_i$. This states that, for each treatment, the success probability is identical for each center. The model satisfies a collapsibility condition for the XY association, because it states that Z is conditionally independent of Y , given X . So, when center effects are negligible and the simpler model fits nearly as well, the estimated treatment effect is approximately the marginal XY odds ratio.

6.4.9 Testing Homogeneity of Odds Ratios

The homogeneous association condition $\theta_{XY(1)} = \dots = \theta_{XY(K)}$ for $2 \times 2 \times K$ tables is equivalent to logistic model (6.9). A test of homogeneous association is implicitly a goodness-of-fit test of this model. The usual G^2 and X^2 test statistics provide this, with $df = K - 1$. They test that the $K - 1$ parameters in the saturated model that are the coefficients of interaction terms [cross-products of the indicator variable for X with $(K - 1)$ indicator variables for categories of Z] all equal 0.

For the eight-center clinical trial data in Table 6.9, $G^2 = 9.75$ and $X^2 = 8.03$ ($df = 7$) do not contradict the hypothesis of equal odds ratios. It is reasonable to summarize the conditional association by a single odds ratio (e.g., $\theta_{\text{MH}} = 2.13$ or $e^{\hat{\beta}} = 2.17$) for all eight partial tables.

6.4.10 Summarizing Heterogeneity in Odds Ratios

In practice, the effect of interest is often similar from stratum to stratum. In multicenter clinical trials comparing a new drug to a standard, for example, if the new drug is truly more beneficial, the population effect is usually positive in each stratum.

In strict terms, however, a model with homogeneous effects is unrealistic. Consider the odds ratio, to illustrate. First, we rarely expect the true odds ratio to be *exactly* the same in each stratum, because of unmeasured covariates that affect it. Breslow (1976) discussed modeling of the log odds ratio using a set of explanatory variables. Second, the model regards the strata effects $\{\beta_k^Z\}$ as fixed effects, treating them as the only strata of interest. Often the strata are merely a sampling of the possible ones. Multicenter clinical trials have data for certain centers but many other centers could have been chosen. Scientists would like their conclusions to apply to all such centers, not only those in the study.

A somewhat different logistic model treats the true log odds ratios in the partial tables as a random sample from a $N(\mu, \sigma^2)$ distribution. Fitting the model yields an estimated mean log odds ratio and an estimated variability about that mean. The inference applies to the population of strata rather than only those sampled. This type of model uses *random effects* in the linear predictor to induce this extra type of variability. In Chapter 13, we discuss GLMs with random effects, and in Section 13.3.5 we fit such a model to [Table 6.9](#).

6.4.11 Propensity Scores in Observational Studies

We finish this section by mentioning a more challenging setting for analyzing conditional associations – observational studies in which we want to compare two groups while controlling for possibly confounding variables x . Rosenbaum and Rubin (1983) proposed methods of adjusting for bias in making such comparisons. They defined the *propensity* as the probability of being in one group, for a given setting of the explanatory variables x . They used logistic regression to estimate how propensity depends on x . In comparing the groups on the response variable, they showed how to control for differing distributions of the groups on x by adjusting for the estimated propensity. This is done by using the propensity to match samples from the groups or to subclassify subjects into several strata consisting of intervals of propensity scores or to adjust directly by entering the propensity in the model.

For any study that is observational rather than randomized, there is still the limitation that propensity score methods, adjust only for observed confounding covariates and not for unobserved ones. Also, the methods work better in larger samples, so observed covariates tend to be more truly balanced in the subclassifications. In various writings, Rubin has pointed out that confidence in causal conclusions based on such methods must rely on how consistent the results are with other evidence and how sensitive the conclusions are to reasonable deviations such as in the effects of unobserved covariates.

6.5 DETECTING AND DEALING WITH INFINITE ESTIMATES

The log-likelihood function for logistic regression models is strictly concave. ML estimates exist and are unique except in certain boundary cases. Estimates do not exist or may be infinite when there is no overlap in the sets of explanatory variable values having $y = 0$ and having $y = 1$ (Albert and Anderson 1984).

6.5.1 Complete or Quasi-complete Separation

The space of explanatory variable values is said to have *complete separation* when a hyperplane can pass through that space such that on one side of that hyperplane $y_i = 0$ for all observations, whereas on the other side, $y_i = 1$ always. This means that there exists a vector \mathbf{b} such that

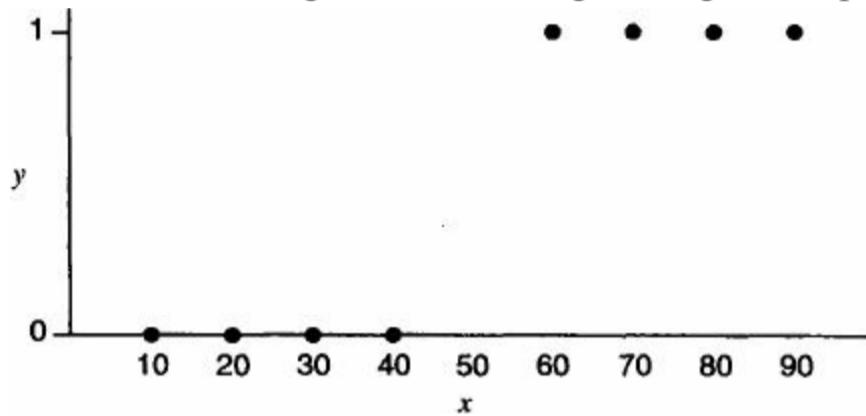
$$\mathbf{b}^T \mathbf{x}_i > 0 \text{ whenever } y_i = 1,$$

$$\mathbf{b}^T \mathbf{x}_i < 0 \text{ whenever } y_i = 0.$$

There is then *perfect discrimination*, as we can predict the sample outcomes perfectly by knowing the predictor values.

[Figure 6.5](#) illustrates for a single explanatory variable. Here, $y = 0$ at $x = 10, 20, 30, 40$, and $y = 1$ at $x = 60, 70, 80, 90$. For $\mathbf{x}_i = (1, x_i)^T$, the predictor $\mathbf{b}^T \mathbf{x}_i = -50 + x_i$ [i.e., $\mathbf{b}^T = (-50, 1)$] gives perfect predictions. An ideal fit has $\hat{\pi} = 0$ for $x < 50$ and $\hat{\pi} = 1$ for $x > 50$. By letting $\hat{\beta} \rightarrow \infty$ and, for fixed $\hat{\beta}$, letting $\hat{\alpha} = -\hat{\beta}(50)$ so that $\hat{\pi} = 0.50$ at $x = 50$, we can generate a sequence with ever-increasing value of the likelihood function that comes successively closer to a perfect fit.

[Figure 6.5](#) Perfect discrimination resulting in an infinite logistic regression parameter estimate.



In practice, most software fails to recognize when some ML estimates are actually infinite. After a few cycles of iterative fitting, the log likelihood looks flat at the working estimate, and convergence criteria are satisfied. Because the log likelihood is so flat and because the variance of $\hat{\beta}_j$ comes from the negative inverse of the matrix of second derivatives, software typically reports huge standard errors. For the data in [Figure 6.5](#), for instance, PROC GENMOD in SAS reports $\text{logit}(\hat{\pi}) = -192.2 + 3.8x$ with standard errors of 8.0×10^8 and 1.5×10^7 .

In practice, an indication of complete separation is when the fitted prediction equation perfectly predicts the response outcome for the entire data, giving $\hat{\pi} = 1.0$ (to many decimal places) whenever $y = 1$ and $\hat{\pi} = 0.0$ whenever $y = 0$. A related indication is that the reported maximized log-likelihood value is 0 to many decimal places. Another warning signal is when standard errors seem unnaturally large. When there is indication of complete separation for a model containing several predictors, using the forward selection algorithm can reveal a subset of them for which complete separation occurs once they are all used.

A weaker condition that causes at least one estimate to be infinite, called *quasi-complete separation*, occurs when a hyperplane separates explanatory variable values with $y = 1$ and with $y = 0$, but cases exist with both outcomes on that hyperplane. For example, this happens if we add to [Figure 6.5](#) two observations at $x = 50$, one with $y = 1$ and one with $y = 0$. With quasi-complete separation, there is not perfect discrimination for all observations. The maximized log likelihood is then strictly less than 0. An indication of quasi-complete separation is that some observations have $\hat{\pi} = 1.0$ or 0.0 . Again, a warning signal is when reported standard errors seem unnaturally large.

When complete or quasi-complete separation do not occur, all ML estimates are finite and unique. Quasi-complete separation is more common than complete separation. It is more liable to happen with qualitative predictors than quantitative predictors. If any category of a qualitative predictor has either no cases with $y = 0$ or no cases with $y = 1$, there is quasi-complete separation when that variable is entered as a factor in the model (i.e., using an indicator variable for that category). With many predictors, it's a good idea to cross-classify each qualitative predictor with y to check for an

empty cell, which is a sufficient condition for quasi-complete separation.

With an infinite estimate, Wald inference is worthless. By contrast, we can still compute likelihood-ratio and score tests and invert them to get a confidence interval. For example, the likelihood still has a maximized value at the infinite estimate for a parameter, so we can compare its value to the value when the parameter is equated to some fixed value such as zero. For the data in [Figure 6.5](#), the likelihood-ratio test statistic for $H_0: \beta = 0$ is 11.09 (df = 1, $P = 0.001$), and the 95% confidence interval for β is $(0.06, \infty)$, so we can conclude that the effect is positive in the population.

6.5.2 Example: Multicenter Clinical Trial with Few Successes

[Table 6.11](#) shows results of a clinical trial conducted at five centers. The purpose was to compare an active drug to placebo for treating fungal infections, with a binary (success, failure) response. For these data, let Y = response, X = treatment (1 = active drug, 0 = placebo), and Z = center.

Table 6.11 Clinical Trial Relating Treatment to Response, Showing also XY and YZ Marginal Tables

Center (Z)	Treatment (X)	Response (Y)		YZ Marginal	
		Success	Failure	Success	Failure
1	Active drug	0	5	0	14
	Placebo	0	9		
2	Active drug	1	12	1	22
	Placebo	0	10		
3	Active drug	0	7	0	12
	Placebo	0	5		
4	Active drug	6	3	8	9
	Placebo	2	6		
5	Active drug	5	9	7	21
	Placebo	2	12		
XY marginal	Active drug	12	36		
	Placebo	4	42		

Source: Data courtesy of Diane Connell, Sandoz Pharmaceuticals Corporation.

Centers 1 and 3 had no successes. Thus, the 5×2 YZ marginal table relating response to center collapsed over treatment, shown on the right side of [Table 6.11](#), contains zero counts. Infinite ML estimates occur for terms in logistic models relating to the YZ association. An example is the model

$$\text{logit}(\pi_{ik}) = \beta x_i + \beta_k^Z.$$

[We take out the intercept from [\(6.9\)](#), so the $\{\beta_k^Z\}$ need no constraint; then, these refer to each center's effect rather than contrasts between each center and a baseline center.] The likelihood function increases continually as β_1^Z and β_3^Z decrease toward $-\infty$; that is, as the logit decreases toward $-\infty$, so the fitted probability of success decreases toward the ML estimate of 0 for those centers.

Because of the infinite estimates, we cannot conduct a Wald test of the center effects in [Table 6.11](#). However, SAS (PROC GENMOD) reports a maximized log-likelihood value of -28.87 for this model and -40.58 when the center term is removed from the model, so the likelihood-ratio statistic for this effect equals 23.42 ($df = 4$).

The counts in the 2×2 marginal table relating response to treatment, shown in the bottom panel of [Table 6.11](#), are all positive. The empty cells affect the center estimates, but not the treatment estimate, for this model. In the limit as the log likelihood increases, the fitted values have a log odds ratio $\hat{\beta} = 1.55$ ($SE = 0.70$). Most software reports this but, instead of $\hat{\beta}_1^Z = \hat{\beta}_3^Z = -\infty$, reports large numbers with extremely large standard errors. For instance, PROC GENMOD in SAS reports values of about -26 for β_1^Z and β_3^Z , with standard errors of about 200,000.

The treatment estimate $\hat{\beta} = 1.55$ also results when we delete centers 1 and 3 from the analysis. When a center contains responses of only one type, it provides no information about this odds ratio. (It does provide information about the size of some other measures, such as the difference of proportions, as discussed above in Section 6.4.6.) Such tables also make no contribution to standard tests of conditional independence, such as the Cochran–Mantel–Haenszel test.

An alternative strategy in multicenter analyses combines centers of a similar type. Then, if each resulting partial table has responses with both outcomes, the inferences use all data. For [Table 6.11](#), perhaps centers 1 and 3 are similar to center 2, since the success rate is very low for that center.

Combining these three centers and refitting the model to this table and the tables for the other two centers yields $\hat{\beta} = 1.56$ ($SE = 0.70$). Usually, this strategy produces results essentially the same as from deleting tables with no outcomes of a particular type.

6.5.3 Remedies When at Least One ML Estimate Is Infinite

What can you do if there is complete or quasi-complete separation and thus at least one ML estimate does not exist? As just mentioned, you can still usually do inference about that effect. For example, you can conduct a likelihood-ratio test. If $\hat{\beta} = \infty$, a profile likelihood confidence interval will have the form (L, ∞) . With quasi-complete separation, some parameter estimates may be unaffected, and their inference will resemble the usual. With small samples and categorical predictors, you can use the specialized exact conditional methods to be presented in Section 7.3.

Alternatively, you can make some adjustment so all estimates are finite. For example, if a category of a qualitative predictor has no cases with $y = 1$, perhaps combine that category with a similar one such that outcomes of both type then occur. Some approaches smooth the data, thus producing finite estimates. The Bayesian approach (Section 7.2) is the best known way of doing that. The amount of smoothing for the resulting estimates depend strongly on the variability in the Bayes prior distribution.

A related way maximizes a *penalized likelihood* function. This adds a term to the ordinary log-likelihood function such that maximizing the amended function smooths the estimates by shrinking them toward 0 (Firth 1993a). Section 7.4.5 introduces this approach, which corresponds to using the Bayesian posterior mode induced by the Jeffreys prior distribution. For the data in [Figure 6.5](#), this method replaces the infinite estimate of β by $\hat{\beta} = 0.067$ ($SE = 0.042$). The corresponding 95% penalized profile likelihood confidence interval is $(0.013, 0.334)$. Its highly asymmetric form about $\hat{\beta}$ reflects the highly nonsymmetric appearance of the log-likelihood function for such data.

6.6 SAMPLE SIZE AND POWER CONSIDERATIONS

In any statistical procedure, the sample size n influences the results. Strong effects are likely to be detected even when n is small. By contrast, detection of weak effects requires large n . A study design should reflect the sample size needed to provide good power for detecting the effect.

6.6.1 Sample Size and Power for Comparing Two Proportions

For test statistics having large-sample normal distributions, power calculations can use ordinary methods. To illustrate, consider a test comparing binomial parameters π_1 and π_2 for two medical treatments. An experiment plans independent samples of size $n_i = n/2$ receiving each treatment. The researchers expect $\pi_i \approx 0.60$ for each, and a difference of at least 0.10 is important. In testing $H_0: \pi_1 = \pi_2$, the variance of $\hat{\pi}_1 - \hat{\pi}_2$ is $\pi_1(1 - \pi_1)/(n/2) + \pi_2(1 - \pi_2)/(n/2) \approx 0.60 \times 0.40 \times (4/n) = 0.96/n$. In particular,

$$z = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - (\pi_1 - \pi_2)}{\sqrt{0.96/n}}$$

has approximately a standard normal distribution for π_1 and π_2 near 0.60.

The power of an α -level test of H_0 is approximately

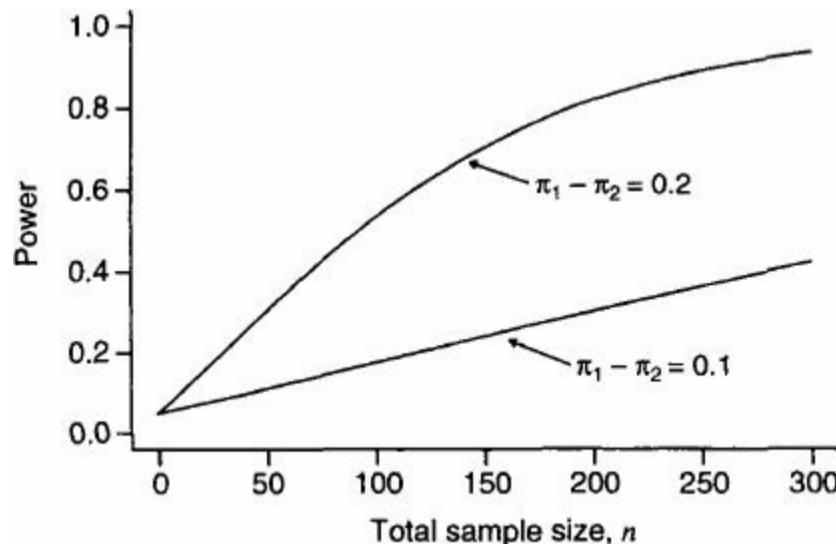
$$P\left[\frac{|\hat{\pi}_1 - \hat{\pi}_2|}{\sqrt{0.96/n}} \geq z_{\alpha/2}\right].$$

When $\pi_1 - \pi_2 = 0.10$, for $\alpha = 0.05$, this equals

$$\begin{aligned} P\left[\frac{(\hat{\pi}_1 - \hat{\pi}_2) - 0.10}{\sqrt{0.96/n}} > 1.96 - 0.10\sqrt{n/0.96}\right] \\ + P\left[\frac{(\hat{\pi}_1 - \hat{\pi}_2) - 0.10}{\sqrt{0.96/n}} < -1.96 - 0.10\sqrt{n/0.96}\right] \\ = P[z > 1.96 - 0.10\sqrt{n/0.96}] + P[z < -1.96 - 0.10\sqrt{n/0.96}] \\ = 1 - \Phi[1.96 - 0.10\sqrt{n/0.96}] + \Phi[-1.96 - 0.10\sqrt{n/0.96}], \end{aligned}$$

where Φ is the standard normal cdf. The power is approximately 0.11 when $n = 50$ and 0.30 when $n = 200$. It is not easy to attain significance when effects are small and the sample size is not very large. [Figure 6.6](#) shows how the power increases in n when $\pi_1 - \pi_2 = 0.10$. By contrast, it also shows how the power improves when $\pi_1 - \pi_2 = 0.20$.

[Figure 6.6](#) Approximate power for testing equality of proportions, with true values near middle of range and $\alpha = 0.05$.



For specified $P(\text{type I error}) = \alpha$ and $P(\text{type II error}) = \beta$ (and hence power = $1 - \beta$), we can determine the sample size needed to attain those values. A study using $n_1 = n_2$ requires approximately

$$n_1 = n_2 = (z_{\alpha/2} + z_{\beta})^2 [\pi_1(1 - \pi_1) + \pi_2(1 - \pi_2)] / (\pi_1 - \pi_2)^2.$$

For a test with $\alpha = 0.05$ and $\beta = 0.10$ when π_1 and π_2 are truly about 0.60 and 0.70, $n_1 = n_2 = 473$. Similarly, with about 473 subjects in each group, a 95% confidence interval has only a 0.10 chance of containing 0 when actually, $\pi_1 = 0.60$ and $\pi_2 = 0.70$.

This sample-size formula is approximate and may underestimate slightly the actual values required. It is adequate for most practical work, though, in which only rough conjectures are available for π_1

and π_2 . Farrington and Manning (1990) and Fleiss et al. (2003, Chap. 4) showed more precise formulas.

6.6.2 Sample Size Determination in Logistic Regression

Consider now the model $\text{logit}[\pi(x_i)] = \alpha + \gamma x_i, i = 1, \dots, n$, in which x is quantitative. [We use γ so as not to confuse with $\beta = P(\text{type II error})$.] The sample size needed to achieve a certain power for testing $H_0: \gamma = 0$ depends on the variance of $\hat{\gamma}$. This depends on $\{\pi(x_i)\}$, and formulas for n use a guess for $\hat{\pi} = \pi(\bar{x})$ and the distribution of X . The effect size is the log odds ratio τ comparing $\pi(\bar{x})$ to $\pi(\bar{x} + s_x)$, the probability at a standard deviation above the mean of x . For a one-sided test when X is approximately normal, Hsieh (1989) derived

$$(6.10) \quad n = [z_\alpha + z_\beta \exp(-\tau^2/4)]^2 (1 + 2\hat{\pi}\delta)/(\hat{\pi}\tau^2),$$

where

$$\delta = [1 + (1 + \tau^2) \exp(5\tau^2/4)]/[1 + \exp(-\tau^2/4)].$$

The value n decreases as $\hat{\pi} \rightarrow 0.50$ and as $|\tau|$ increases.

We illustrate for modeling the effect of $x = \text{cholesterol level}$ on the probability of severe heart disease for a population for which that probability at an average level of cholesterol is about 0.08. Researchers want the test to be sensitive to a 50% increase in this probability, for a standard deviation increase in cholesterol. The odds of severe heart disease at the mean cholesterol level equal $0.08/0.92 = 0.087$, and the odds one standard deviation above the mean equal $0.12/0.88 = 0.136$. The odds ratio equals $0.136/0.087 = 1.57$, and $\tau = \log(1.57) = 0.450$. For $\alpha = 0.05$ and $\beta = 0.10$, $\delta = 1.306$ and $n = 612$.

6.6.3 Sample Size in Multiple Logistic Regression

A multiple logistic regression model requires larger n to detect effects. Let R denote the multiple correlation between the predictor X of interest and the others in the model. The formula (6.10) for n divides by $(1 - R^2)$. In that formula, $\hat{\pi}$ is evaluated at the mean of all the explanatory variables, and the odds ratio refers to the effect of X at the mean level of the other predictors.

Consider the example in Section 6.2.2 when blood pressure is also a predictor. If the correlation between cholesterol and blood pressure is 0.40, we need $n \approx 612/[1 - (0.40)^2] = 729$.

These formulas provide, at best, very approximate indications of sample size. Most applications have only a crude guess for $\hat{\pi}$ and R , and X may be far from normally distributed.

6.6.4 Power for Chi-Squared Tests in Contingency Tables

When hypotheses are false, squared normal and X^2 and G^2 statistics have large-sample noncentral chi-squared distributions (Section 5.3.8). Suppose that H_0 is equivalent to model M for a contingency table. Let π_i denote the true probability in cell i , and let $\pi_i(M)$ denote the value to which the ML estimate $\hat{\pi}_i$ for model M converges, where $\sum_i \pi_i = \sum_i \pi_i(M) = 1$. For a multinomial sample of size n , the noncentrality parameter for X^2 equals

$$(6.11) \quad \lambda = n \sum_i \frac{[\pi_i - \pi_i(M)]^2}{\pi_i(M)}.$$

This has the same form as X^2 , with π_i in place of the sample proportion p_i , and $\pi_i(M)$ in place of $\hat{\pi}_i$. The noncentrality parameter for G^2 equals

$$(6.12) \quad \lambda = 2n \sum_i \pi_i \log \frac{\pi_i}{\pi_i(M)}.$$

When H_0 is true, all $\pi_i = \pi_i(M)$. Then, for either statistic, $\lambda = 0$ and the ordinary (central) chi-squared distribution applies.

To determine the approximate power for a chi-squared test with $df = v$, (1) choose a hypothetical set of true values $\{\pi_i\}$, (2) calculate $\{\pi_i(M)\}$ by fitting to $\{\pi_i\}$ the model M for H_0 , (3) calculate the noncentrality parameter λ , and (4) calculate $P[X_{v,\lambda}^2 > \chi_v^2(\alpha)]$. [Table 6.12](#) shows an excerpt from a table of noncentral chi-squared probabilities for step 4 with $\alpha = 0.05$.

Table 6.12 Power of Chi-Squared Test for $\alpha = 0.05$

df	Noncentrality													
	0.0	0.2	0.4	0.6	0.8	1.0	2.0	3.0	4.0	5.0	7.0	10.0	15.0	25.0
1	.050	.073	.097	.121	.146	.170	.293	.410	.516	.609	.754	.885	.972	.998
2	.050	.065	.081	.098	.115	.133	.226	.322	.415	.504	.655	.815	.944	.996
3	.050	.062	.075	.088	.102	.116	.192	.275	.358	.440	.590	.761	.917	.993
4	.050	.060	.071	.082	.093	.106	.172	.244	.320	.396	.540	.716	.891	.989
6	.050	.058	.066	.075	.084	.094	.146	.206	.270	.336	.468	.644	.843	.980
8	.050	.057	.064	.071	.079	.087	.131	.182	.238	.296	.417	.588	.799	.968
10	.050	.056	.062	.068	.075	.082	.121	.166	.215	.268	.379	.542	.760	.956
20	.050	.053	.056	.060	.063	.066	.096	.125	.158	.193	.273	.402	.611	.883
50	.050	.052	.054	.056	.059	.061	.076	.092	.110	.129	.173	.250	.398	.687

Source: Reprinted with permission from G. E. Haynam, Z. Govindarajulu, and F. C. Leone, in *Selected Tables in Mathematical Statistics*, eds. H. L. Harter and D. B. Owen. Chicago: Markham, 1970.

6.6.5 Power for Testing Conditional Independence

We use an example based on one in O'Brien (1986). A standard fetal heart rate monitoring test predicts whether a fetus will require nonroutine care following delivery. The standard test has categories (worrisome, reassuring). The response Y is whether the newborn required some nonroutine medical care during the first week after birth ($1 = \text{yes}$, $0 = \text{no}$). A new fetal heart rate monitoring test is developed, having categories (very worrisome, somewhat worrisome, reassuring). A physician plans to study whether this new test can help make predictions about the outcome; that is, given the result of the standard test, is there an association between the response and the result of the new test? A relevant statistic tests the effect of the new monitoring test in the logistic model having the new test (N) and the standard test (S) as qualitative predictors.

To help select n , a statistician asks the physician to conjecture about the joint distribution of the explanatory variables, with questions such as “What proportion of the cases do you think will be scored ‘reassuring’ by both tests?” For each NS combination, the physician also guessed $P(Y = 1)$. [Table 6.13](#) shows one scenario for marginal and conditional probabilities. These yield a joint distribution $\{\pi_{ijk}\}$ from their product, such as $0.04 \times 0.40 = 0.016$ for the proportion of cases judged worrisome by the standard test and very worrisome by the new test and requiring nonroutine medical care. These joint probabilities yield fitted probabilities $\pi(M_0)$ and $\pi(M_1)$ for the null and alternative logit models. (We can get these by entering $\{\pi_{ijk}\}$ in percentage form as counts in software for logistic regression, fitting the relevant model, and dividing the fitted counts by 100 to get the fitted joint probabilities.) The likelihood-ratio test comparing these models has noncentrality [\(6.12\)](#) with $\pi(M_1)$ playing the role of π and $\pi(M_0)$ playing the role of $\pi(M)$.

[Table 6.13](#) Scenario for Power Computation

Standard Test	New Test	Joint Probability	$P(\text{nonroutine care})$
Worrisome	Very worrisome	0.04	0.40
	Somewhat worrisome	0.08	0.32
	Reassuring	0.04	0.27
Reassuring	Very worrisome	0.02	0.30
	Somewhat worrisome	0.18	0.22
	Reassuring	0.64	0.15

Source: Reprinted with permission from O'Brien (1986).

For the scenario in [Table 6.13](#), the noncentrality equals $0.00816n$, with $\text{df} = 2$. For $n = 400$, 600 , and 1000 , the approximate powers when $\alpha = 0.05$ are 0.35 , 0.49 , and 0.73 . This scenario predicts 64% of the observations to occur at only one combination of the factors. The lack of dispersion for the factors weakens the power.

6.6.6 Effects of Sample Size on Model Selection and Inference

The effects of sample size suggest some cautions for model selection. For small n , the most parsimonious model accepted in a goodness-of-fit test may be quite simple. By contrast, larger samples usually require more complex models to pass goodness-of-fit tests. Then, some effects that are statistically significant may be weak and substantively unimportant. With large n it may be adequate to use a model that is simpler than models that pass goodness-of-fit tests. An analysis that focuses solely on goodness-of-fit tests is incomplete. It is also necessary to estimate model parameters and describe strengths of effects.

These remarks merely reflect limitations of significance testing. In many areas of application, null hypotheses are rarely true. With large enough n , they will be rejected. A more relevant concern is whether the difference between true parameter values and null hypothesis values is sufficient to be important. Many methodologists overemphasize testing and underutilize estimation methods such as confidence intervals. When the P -value is small, a confidence interval specifies the extent to which H_0 may be false, thus helping us determine whether rejecting it has practical importance. When the P -value is not small, the confidence interval indicates whether some plausible parameter values are far from H_0 . A wide confidence interval containing the H_0 value indicates that the test had weak power at important alternatives.

NOTES

Section 6.1: Strategies in Model Selection

6.1 AIC, BIC: For cogent arguments supporting the use of AIC, see Burnham and Anderson (2010). A modified version is recommended if the number of parameters is large. Some statisticians believe that BIC can select an overly simple model. For this and other critiques, see articles by Gelman and Rubin, Firth and Kuha, Raftery, Weakliem, and Xie, in the February 1999 issue of *Sociological Methods and Research*.

Section 6.2: Logistic Regression Diagnostics

6.2 Diagnostics: Olive and Hawkins (2005) presented graphics that are useful for variable selection. As an alternative to the residual methods discussed, smoothing the residuals before plotting them (e.g., using methods to be presented in Section 7.4) can be helpful (Fowlkes 1987, Lloyd 1999, Sec. 5.4). Cook and Weisberg (1999, Chap. 22) and Landwehr et al. (1984) showed other examples of useful diagnostic plots. For other logistic regression diagnostics, see Copas (1988), who also considered resistant fitting methods (e.g., to take misclassification into account), Hosmer and Lemeshow (2000, Chap. 5), Johnson (1985), and Pregibon (1981).

Section 6.3: Summarizing the Predictive Power of a Model

6.3 R^2 measures: Amemiya (1981), Efron (1978), Hu et al. (2005), Liao and McGee (2003), Maddala (1983), Schemper (2003), and Zheng and Agresti (2000) and references therein reviewed R^2 measures for binary regression. Hosmer and Lemeshow (2000, Sec. 5.2.3) discussed classification tables and their limitations. Pepe (2004) and references therein surveyed ROC methodology.

Section 6.4: Mantel–Haenszel and Related Methods for Multiple 2 × 2 Tables

6.4 DIF: One application of CMH methods is *differential item functioning*: comparing groups in terms of how different they are in responding to items on a questionnaire, after adjusting for overall abilities or scores. See Holland and Wainer (1993).

6.5 Breslow–Day test: An analog of ϕ_{MH} and δ_{MH} summarizes relative risks from several strata (Greenland and Robins 1985). Breslow and Day (1980, p. 142) proposed an alternative large-sample test of homogeneity of odds ratios. In each partial table let $\{\hat{\mu}_{ijk}\}$ have the same marginals as the data observed, yet have odds ratio equal to ϕ_{MH} . Their test statistic has the Pearson form comparing $\{n_{ijk}\}$ to $\{\hat{\mu}_{ijk}\}$. Tarone (1985) showed that, because of the inefficiency of ϕ_{MH} , the Breslow–Day statistic must be adjusted for it to have exactly a limiting chi-squared null distribution with $df = K - 1$. This adjustment is usually minor. Other work on comparing odds ratios and estimating a common value includes Breslow and Day (1980, Sec. 4.4), Donner and Hauck (1986), Gart (1970), Jones et al. (1989), and Liang and Self (1985). For modeling the odds ratio, see Breslow (1976), Breslow and Day (1980, Sec. 7.5), and Prentice (1976a). Breslow emphasized retrospective studies, in which the conditional approach is natural since the outcome totals are fixed.

Section 6.5: Detecting and Dealing with Infinite Estimates

6.6 Infinite ML: For discussion of this topic, including other link functions and GLMs, see Albert and Anderson (1984), Haberman (1974a), Santner and Duffy (1986), Silvapulle (1981), and Wedderburn (1976).

6.7 High imbalance: King and Zeng (2001) and Owen (2007) discussed applications in which

one outcome category is much more common than the other. Examples include rare diseases, fraudulent use of a credit card, and non-spam email messages in spam folders. King and Zeng proposed a sampling design of sampling all possible cases of the rare outcome and a much smaller fraction of the other outcome. Owen showed that under a sampling scheme for which $n \rightarrow \infty$ while the number of outcomes in one category remains finite, a limit exists for the estimated parameter vector that depends on the distribution of the x values.

Section 6.6: Sample Size and Power Considerations

6.8 Noncentral chi-squared: Gail and Gart (1973) and Suissa and Shuster (1985) studied sample size for obtaining fixed power in Fisher's test. Farrington and Manning (1990) considered sample size for nonnull effects for the difference of proportions and relative risk using score-type tests. For sample size determination in logistic regression, see Hsieh et al. (1998), Lyles et al. (2006), Schoenfeld and Borenstein (2005), Vaeth and Skovlund (2004), and Whittemore (1981). Lachin (1977) considered $I \times J$ tables. Drost et al. (1989), Haberman (1974a, pp. 109-112), Meng and Chapman (1966), Mitra (1958), and Patnaik (1949) derived theory for asymptotic nonnull behavior of chi-squared statistics; see also Section 16.3.5. O'Brien's (1986) simulation results suggested that the noncentral chi-squared approximation for G^2 holds well for a wide range of powers. Read and Cressie (1988, pp. 147-148) listed other articles that studied the nonnull behavior of X^2 and G^2 .

EXERCISES

Applications

6.1 For the horseshoe crab mating data, the maximized log-likelihood value is -112.88 for the model with only an intercept, -97.87 for the model with weight as a predictor, -97.23 for the model with width as a predictor, and -96.45 for the model using both as predictors. Conduct **(a)** a test of $H_0: \beta_1 = \beta_2 = 0$ for the joint effects, and **(b)** separate tests for the partial effects. Why does neither test in part (b) show evidence of an effect when the test in part (a) shows strong evidence?

6.2 For the horseshoe crab mating data. [Table 6.14](#) shows ML estimates for two models using weight and color (with dark color as the baseline) as predictors of satellite presence. Compare the models using a likelihood-ratio test and using AIC. Select a model, and interpret its estimates.

[Table 6.14](#) Effects for Two Models with Predictors of Crab Satellites, for Exercise 6.2

Term	Model 1		Model 2	
	Estimate	SE	Estimate	SE
Intercept	-4.53	1.00	-1.19	2.30
Weight	1.69	0.39	0.19	1.03
Color 1	1.27	0.85	-0.43	5.40
Color 2	1.41	0.54	-1.27	2.58
Color 3	1.08	0.59	-6.73	3.44
Weight \times Color 1			0.85	2.16
Weight \times Color 2			1.21	1.14
Weight \times Color 3			3.56	1.56
Log-likelihood	-94.27		-90.83	
AIC	198.54		197.66	

6.3 The book's website (www.stat.ufl.edu/~aa/cda/cda.html) has a $2 \times 3 \times 2 \times 2$ table relating responses on frequency of attending religious services, political views, opinion on making birth control available to teenagers, and opinion about whether premarital sex before marriage is wrong. Treating opinion about premarital sex as the response variable, use backward elimination to select a model. Interpret.

6.4 For [Table 10.1](#), treating marijuana use as the response variable, build a model with alcohol use, cigarette use, gender, and race as potential explanatory variables. Summarize your strategy for selecting a model, and interpret your final choice of model.

6.5 For [Table 6.4](#), fit the stage 3 model denoted there by $(E * P + G)$. Use parameter estimates to interpret the G effect and the dependence of the E effect on P .

6.6 According to the *Independent* newspaper (London, Mar. 8, 1994), the Metropolitan Police in London reported 30,475 people as missing in the year ending March 1993. For those of age 13 or less, 33 of 3271 missing males and 38 of 2486 missing females were still missing a year later. For ages 14 to 18, the values were 63 of 7256 males and 108 of 8877 females; for ages 19 and above, the values were 157 of 5065 males and 159 of 3520 females. Analyze by building a model, and interpret. (Thanks to Pat Altham for showing me these data.)

6.7 Fowlkes et al. (1988) reported results of a survey of employees of a large national corporation to determine how satisfaction depends on race, gender, age, and regional location. The data are at the book's website. Build a logistic model for these data and carefully interpret the parameter estimates.

6.8 [Table 6.15](#) shows the results of a study about Y = whether a patient having surgery with general anesthesia experienced a sore throat on waking (0 = no, 1 = yes) as a function of the D = duration of the surgery (in minutes) and the T = type of device used to secure the airway (0 = laryngeal mask airway, 1 = tracheal tube). Use a model-building strategy to select a logistic model for these predictors. For your model, interpret parameter estimates, and conduct inference

about the effects.

Table 6.15 Data for Exercise 6.8 on Surgery and Sore Throats

Patient	D	T	Y	Patient	D	T	Y	Patient	D	T	Y
1	45	0	0	13	50	1	0	25	20	1	0
2	15	0	0	14	75	1	1	26	45	0	1
3	40	0	1	15	30	0	0	27	15	1	0
4	83	1	1	16	25	0	1	28	25	0	1
5	90	1	1	17	20	1	0	29	15	1	0
6	25	1	1	18	60	1	1	30	30	0	1
7	35	0	1	19	70	1	1	31	40	0	1
8	65	0	1	20	30	0	1	32	15	1	0
9	95	0	1	21	60	0	1	33	135	1	1
10	35	0	1	22	61	0	0	34	20	1	0
11	75	0	1	23	65	0	1	35	40	1	0
12	45	1	1	24	15	1	0				

Source: Data from "Binary Data" by D. Collett, in *Encyclopedia of Biostatistics*, 2nd ed. Hoboken, NJ: Wiley, 2005, pp. 439–446.

6.9 Refer to the previous exercise. Use a measure of predictive power to compare the fits of various models to these data.

6.10 Refer to the previous two exercises. For your preferred model:

- a. Summarize predictive power using classification tables with $\pi_0 = 0.50$ and $\pi_0 = \bar{y}$. In each case, report and interpret the sensitivity and specificity.
- b. Summarize predictive power using an ROC curve. Report and interpret the concordance index.

6.11 Discern the reasons that Simpson's paradox occurs for the graduate admissions data of [Table 6.6](#).

6.12 Refer to Exercise 2.15 on graduate school admissions and gender. Fit the model of no G effect, given the department. Use X^2 to test the fit. Obtain standardized residuals, explain how they relate to X^2 , and interpret the lack of fit.

6.13 Conduct a residual analysis for the independence model with [Table 5.5](#) on treating leprosy. What type of lack of fit is indicated?

6.14 For the horseshoe crab data, use methods such as Section 6.3 shows to evaluate predictive power for logistic models that include weight and color as explanatory variables.

6.15 [Table 6.16](#) refers to the effectiveness of immediately injected or $1\frac{1}{2}$ -hour-delayed penicillin in protecting rabbits against lethal injection with β -hemolytic streptococci.

Table 6.16 Data for Exercise 6.15 on Penicillin Treatment for Streptococcus

Penicillin Level	Delay	Response	
		Cured	Died
$\frac{1}{8}$	None	0	6
	$1\frac{1}{2}$ h	0	5
$\frac{1}{4}$	None	3	3
	$1\frac{1}{2}$ h	0	6
$\frac{1}{2}$	None	6	0
	$1\frac{1}{2}$ h	2	4
1	None	5	1
	$1\frac{1}{2}$ h	6	0
4	None	2	0
	$1\frac{1}{2}$ h	5	0

Source: Reprinted with permission from Mantel (1963).

- a. Let $X = \text{delay}$, $Y = \text{whether cured}$, and $Z = \text{penicillin level}$. Fit the logistic model (6.4). Argue that the pattern of 0 cell counts suggests that (with no intercept) $\hat{\beta}_1^Z = -\infty$ and $\hat{\beta}_5^Z = \infty$. What does your software report?
- b. Using the logistic model, conduct the likelihood-ratio test of XY conditional independence. Interpret.

- c. Test XY conditional independence using the Cochran–Mantel–Haenszel test. Interpret.
- d. Estimate the XY conditional odds ratio using (i) ML with the logistic model, and (ii) the Mantel–Haenszel estimate. Interpret.

6.16 Refer to [Table 2.6](#). Use the CMH statistic to test independence of death penalty verdict and victim's race, controlling for defendant's race. Conduct another test of this hypothesis, and compare results.

6.17 Treatments A and B were compared on a binary response for 40 pairs of subjects matched on relevant covariates. For each pair, treatments were assigned to the subjects randomly. Twenty pairs of subjects made the same response for each treatment. Six pairs had a success for the subject receiving A and a failure for the subject receiving B, whereas the other 14 pairs had a success for B and a failure for A. Use the Cochran–Mantel–Haenszel procedure to test independence of response and treatment. (In Section 11.1 we present an equivalent test, McNemar's test.)

6.18 For the data summarized in Figure 1 of the 2011 *Lancet* article by Rothwell et al. (377: 31–41) from eight studies on the effect of daily aspirin on cancer deaths, conduct a meta-analysis that combines a significance test with a confidence interval to summarize the size of effect. Interpret.

6.19 For the data summarized in Figure 1 of the 2010 *American Statistician* article by Kulinskaya et al. (64: 350–356), conduct a meta-analysis that combines a significance test with a confidence interval to summarize the size of effect. Interpret.

6.20 A data set at the text website from a 2005 article by D. Potter (*Statist. Med.* 24: 693–708) describes results from a study in which subjects received a drug and the outcome measures whether the subject became incontinent ($y = 1$, yes; $y = 0$, no). The three explanatory variables are lower urinary tract variables that represent drug-induced physiological changes.

- a. Find the prediction equations when each predictor is used separately in logistic regressions.
- b. Try to fit a main-effects logistic model containing all three predictors. What does your software report for the effects and their standard errors? (The ML estimates are actually $-\infty$ for x_1 and x_2 and ∞ for x_3 .) Can you see a pattern in the data that is responsible for this behavior?

6.21 Refer to the example of complete separation in Section 6.5.1. For the 8 observations, randomly generate values for a second predictor from the $N(0, 1)$ distribution. Taking both explanatory variables in your model, is there still complete separation? Is there quasi-complete separation? What does your software report for the model parameter estimates and SE values?

6.22 Refer to the multicenter clinical trial of [Table 6.11](#).

- a. Fit the main effects model considered in the text with your favorite software (omitting the intercept), and summarize results.
- b. For Center 1, add ε successes for the active treatment, and report the impact (if any) on β_1^Z and β_2^Z . Do this for $\varepsilon = 10^{-6}$, $\varepsilon = 10^{-3}$, $\varepsilon = 0.50$. Do such centers give any information about the treatment log odds ratio effect, as described by β and its SE ?

6.23 Apply the logistic regression model to the 2×2 table consisting of the data for Center 5 in [Table 6.9](#), where $x = 1$ for drug and $x = 0$ for control.

- a. Report the ML estimate $\hat{\beta}$.
- b. What does your software report when you try to fit this model? Explain why.
- c. Can you construct a 95% confidence interval for β ? Show how.

6.24 For the example in Section 6.6.1, suppose $\pi_1 = 0.70$ and $\pi_2 = 0.60$. What sample size is needed for the test to have approximate power 0.80, when $\alpha = 0.05$, for (a) $H_a: \pi_1 \neq \pi_2$ and (b) $H_a: \pi_1 > \pi_2$?

6.25 For the example in Section 6.6.1 with equal treatment sample sizes, suppose $\pi_1 = 0.63$ and

$\pi_2 = 0.57$. Explain why the joint probabilities in the 2×2 table are 0.315 and 0.185 for treatment A and 0.285 and 0.215 for treatment B. For the model of independence, explain why the fitted joint probabilities are 0.30 for success and 0.20 for failure, in each row. Show that X^2 has noncentrality parameter $0.00375n$ and $df = 1$. For $n = 200$ and $\alpha = 0.05$, find the power.

6.26 An experiment is designed to compare two treatments on a three-category response. The researcher expects the conditional distributions to be approximately $(0.2, 0.2, 0.6)$ and $(0.3, 0.3, 0.4)$.

a. With 100 observations for each treatment and $\alpha = 0.05$, find the approximate power to compare the distributions using (i) X^2 and (ii) G^2 . Compare results.

b. What sample size is needed for each treatment for the tests in (a) to have approximate power 0.90?

6.27 The horseshoe crab width values in [Table 4.3](#) have $\bar{x} = 26.3$ and $s_x = 2.1$. If the true relationship were similar to the fitted equation in Section 5.1.3, about how large a sample yields $P(\text{type II error}) = 0.10$, with $\alpha = 0.05$, for testing $H_0: \beta = 0$ against $H_a: \beta > 0$?

6.28 This book's website (www.stat.ufl.edu/~aa/cda/cda.html) contains a five-way table relating occupational aspirations (high, low) to gender, residence, IQ, and socioeconomic status. Analyze these data.

6.29 In recent years there has been controversy about the effects of rosiglitazone (an antidiabetic drug) on myocardial infarction (MI) and cardiovascular mortality. Review the 2010 meta-analysis by S. Nissen and K. Wolski in *Archives of Internal Medicine* (14:1191–1201). Conduct your own analysis of the effects of rosiglitazone on MI.

Theory and Methods

6.30 For a sequence of s nested models M_1, \dots, M_s , model M_s is the most complex. Let v denote the difference in residual df between M_1 and M_s .

a. Explain why for $j < k$, $G^2(M_j|M_k) \leq G^2(M_j|M_s)$.

b. Assume model M_j , so that M_k also holds when $k > j$. For all $k > j$, as $n \rightarrow \infty$, $P[G^2(M_j|M_k) > \chi_v^2(\alpha)] \leq \alpha$. Explain why.

c. Gabriel (1966) suggested a simultaneous testing procedure in which, for each pair of models, the critical value for differences between G^2 values is $\chi_v^2(\alpha)$. The final model accepted must be more complex than any model rejected in a pairwise comparison. Since part (b) is true for all $j < k$, argue that Gabriel's procedure has type I error probability no greater than α .

6.31 Prove that the Pearson residuals for the linear logit model applied to a $I \times 2$ contingency table satisfy $X^2 = \sum_{i=1}^I e_i^2$. [Hint: Start with the X^2 sum over the $2I$ cells and combine the two terms from the same row.] Note that this holds for a binomial GLM with a linear trend for *any* link function.

6.32 For ungrouped binary data, explain why when π_i is near 1, residuals are necessarily either small and positive or large and negative. What happens when π_i is near 0?

6.33 For a $2 \times 2 \times K$ table from a multicenter clinical trial, one center has entries $(0, n)$ in row 1 and $(0, 2n)$ in row 2 (i.e., no successes for either treatment).

a. Explain why there is *no* information in this table about whether there is an association, regardless of the value of n . [Hint: Show that $\hat{\pi}_1 - \hat{\pi}_2 = 0$ has estimated null $SE = 0$, and the P -value is 1.0 for Fisher's exact test or for an unconditional exact test.]

b. Explain why there *is* information in the table about the size of association, in terms of the difference of proportions, and the precision of information increases as n increases. Illustrate by finding the 95% score confidence intervals for π_1 , π_2 , and $\pi_1 - \pi_2$, when $n = 10$ and when $n = 100$. (See www.stat.ufl.edu/~aa/cda/R for R functions. Note that Wald intervals are

useless for such data.)

6.34 Refer to logit model (6.4) for a $2 \times 2 \times K$ contingency table $\{n_{ijk}\}$. Using a basic result for testing in exponential families, explain why uniformly most powerful unbiased tests of conditional XY independence are based on $\sum_k n_{11k}$ (Birch 1964b; Lehmann and Romano 2005, Sec. 4.8).

6.35 Suppose that $\{\pi_{ijk}\}$ in a $2 \times 2 \times 2$ table are, by row, $(0.15, 0.10 / 0.10, 0.15)$ when $Z = 1$ and $(0.10, 0.15/0.15, 0.10)$ when $Z = 2$. For testing conditional XY independence with logistic models having Y as a response, explain why the likelihood-ratio test comparing models $X + Z$ and Z is not consistent but the likelihood-ratio test of fit of the XY conditional independence model is.

6.36 For 2×2 tables with all marginal totals positive, explain what patterns of 0 cell counts correspond to (a) complete separation and (b) quasi-complete separation.

6.37 For k explanatory variables, suppose logistic regression has finite parameter estimates when used with each predictor alone. Explain why infinite estimates could occur when the predictors are all used in a main-effects model. Sketch a graph with $k = 2$ to illustrate this.

6.38 In Table 6.11, suppose the outcome of 0 successes for the active drug in Centers 1 and 3 was instead a positive count, but there were still no successes for placebo in those centers. Explain why all estimates would be finite for the main-effects model fitted in Section 6.5.2, but infinite estimates would occur for the more general model permitting center-by-treatment interaction.

6.39 Explain why complete or quasi-complete separation would not cause ML estimates to be infinite if you were using the identity link function but might cause other problems with the iterative fitting process.

6.40 For a GLM, let $\hat{\mu}^{(-)} = (\hat{\mu}^{(-1)}, \dots, \hat{\mu}^{(-n)})$, where $\hat{\mu}^{(-i)}$ denotes the estimate of $E(Y_i)$ for observation i after fitting the model without that observation. The *leave-one-out cross-validation* adjustment to the predictive measure $R(\mathbf{y}, \hat{\mu})$ is $\text{corr}(\mathbf{y}, \hat{\mu}^{(-)})$. For binary data, consider the model, $\text{logit}(\pi_i) = \alpha$ for all i . Show that $\hat{\pi}_i = \bar{y}$, $\hat{\pi}^{(-i)} = [n/(n-1)][\bar{y} - (1/n)y_i]$, and hence $\text{corr}(\mathbf{y}, \hat{\pi}^{(-)}) = -1$. This suggests that leave-one-out cross-validation can be misleading for estimating the correlation with model $\text{logit}(\pi_i) = \alpha + \beta x$ when the true effect is very weak (Zheng and Agresti 2000).

6.41 Using graphs or tables, explain what is meant by *no interaction* in modeling response variable Y and explanatory variables X and Z when:

- a. All variables are continuous (multiple regression).
- b. Y and X are continuous, Z is categorical (analysis of covariance).
- c. Y is continuous, X and Z are categorical (two-way ANOVA).
- d. Y is binary, X and Z are categorical (logistic regression).

¹We discussed the parsimony issue, with examples, in Sections 3.3.8.5.2.2, and 5.3.10.

CHAPTER 7

Alternative Modeling of Binary Response Data

In Chapters 5 and 6 we have focused on logistic regression modeling of binary response data. This chapter presents some alternative ways of modeling binary data.

Although the logit is the most popular link function for binary responses, other links are sometimes more appropriate. In Section 7.1 we present the *probit model*, which results from normal latent variable models. We also present models using a double log link function, which imply nonsymmetric response curves. In Section 7.2 we introduce Bayesian approaches for modeling binary responses. For small samples or models with many parameters, ordinary ML inference may perform poorly. In Section 7.3 we discuss *conditional logistic regression*. This method uses conditioning arguments to eliminate nuisance parameters and can provide inference based on exact distributions rather than large-sample approximations.

In Section 7.4 we present methods for discovering structure by *smoothing* the data. A simple version of *kernel smoothing* estimates a probability at any point simply by averaging binary data at nearby points. The *penalized likelihood* method maximizes an adjusted (“penalized”) version of the likelihood function, producing parameter estimates that tend to be more smooth, with some of the estimates possibly even shrinking to 0 under one type of penalty. The *generalized additive model* extends generalized linear models by allowing an unspecified function of an explanatory variable as a predictor. The final section discusses some issues that arise in using the models for binary data sets having very large numbers of potential explanatory variables.

7.1 PROBIT AND COMPLEMENTARY LOG–LOG MODELS

In this section we present two alternatives to logistic models for binary responses. These models for $\pi(x) = P(Y=1)$ have form

$$(7.1) \quad g[\pi(\mathbf{x})] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

for a link function g other than the logit.

7.1.1 Probit Models: Three Latent Variable Motivations

In Section 4.2.6 we saw that in toxicology studies with dosage predictor x , a latent variable model naturally leads to a binary regression model. Specifically, a *tolerance distribution* with cdf F for the dosage that induces a success response implies a model in which the link g is the inverse of a standardized cdf Φ for the family to which F belongs. That is, the model has form

$$(7.2) \quad \Phi^{-1}[\pi(x)] = \alpha + \beta x,$$

with link function Φ^{-1} . Toxicological experiments often measure dosage as the log concentration and take the tolerance distribution to be approximately $N(\mu, \sigma^2)$ for unknown μ and σ (Bliss 1935). If F is a normal cdf, then $\pi(x)$ satisfies this model with Φ as the standard normal cdf. It is called the *probit model*. The probit link function is $\Phi^{-1}(\cdot)$.

A related normal latent variable model, referred to as a *threshold model*, also implies the probit model. This model assumes there is an unobserved continuous response y^* such that the observed response $y = 0$ if $y^* \leq \tau$ and $y = 1$ if $y^* > \tau$. Suppose that $y^* = \mu + \epsilon$, where $\mu = \alpha + \beta x$ and where $\{\epsilon_i\}$ are independent from a $N(0, \sigma^2)$ distribution. Then,

$$\begin{aligned} P(Y = 1) &= P(Y^* > \tau) = P(\alpha + \beta x + \epsilon > \tau) \\ &= P(-\epsilon < \alpha + \beta x - \tau) = \Phi[(\alpha + \beta x - \tau)/\sigma]. \end{aligned}$$

(Note that $-\epsilon$ has the same distribution as ϵ .) There is no information in the data about σ or the threshold τ . An equivalent model results if we multiply $(\alpha, \beta, \sigma, \tau)$ by any positive constant. For identifiability, we set $\sigma = 1$ and $\tau = 0$. Thus, the probit model results. The logistic model follows when ϵ has instead a standard logistic distribution.

A third normal latent variable derivation of the probit model is based on utility functions. Consider the choice between two options, such as two product brands. Let U_0 denote the *utility* of outcome $y = 0$ and U_1 the utility of $y = 1$. For $y = 0$ and 1 , suppose that $U_y = \alpha_y + \beta_y x + \epsilon_y$. A particular subject selects $y = 1$ if their $U_1 > U_0$. Now suppose that ϵ_0 and ϵ_1 are independent $N(0, 1)$ random variables. Then,

$$\begin{aligned} P(Y = 1) &= P(\alpha_1 + \beta_1 x_1 + \epsilon_1 > \alpha_0 + \beta_0 x_0 + \epsilon_0) \\ &= P\{(\epsilon_0 - \epsilon_1)/\sqrt{2} < [(\alpha_1 - \alpha_0) + (\beta_1 - \beta_0)x]/\sqrt{2}\} = \Phi(\alpha^* + \beta^* x), \end{aligned}$$

where $\alpha^* = (\alpha_1 - \alpha_0)/\sqrt{2}$ and $\beta^* = (\beta_1 - \beta_0)/\sqrt{2}$. This is the probit model.

All three of these latent variable approaches extend directly to multiple explanatory variables. The probit extends to an inverse t link, for which corresponding latent variable models can better accommodate outliers.

7.1.2 Probit Models: Interpreting Effects

For the probit model with a single quantitative predictor, the response curve for $\pi(x)$ [or for $1 - \pi(x)$, when $\beta < 0$] has the appearance of the normal cdf with mean $\mu = -\alpha/\beta$ and standard deviation $\sigma = 1/|\beta|$. Since 68% of the normal density falls within a standard deviation of the mean, $1/|\beta|$ is the distance between x values where $\pi(x) = 0.16$ or 0.84 and where $\pi(x) = 0.50$. The instantaneous rate of change in $\pi(x)$ is $\partial\pi(x)/\partial x = \beta\phi(\alpha + \beta x)$, where $\phi(\cdot)$ is the standard normal density function. The rate is highest when $\alpha + \beta x = 0$ (i.e., at $x = -\alpha/\beta$), where it equals $\beta/(2\pi)^{1/2} = 0.40\beta$ (for $\pi = 3.14\dots$). At that point, $\pi(x) = \frac{1}{2}$.

By comparison, in logistic regression with parameter β , the curve for $\pi(x)$ is a logistic cdf with standard deviation $\pi/|\beta|\sqrt{3}$. Its rate of change in $\pi(x)$ at $x = -\alpha/\beta$ is 0.25β . The rates of change where $\pi(x) = \frac{1}{2}$ are the same for the cdf's corresponding to the probit and logistic curves when the logistic β is $0.40/0.25 = 1.60$ times the probit β . The standard deviations are the same when the logistic β is $\pi/\sqrt{3} = 1.81$ times the probit β . When both models fit well, parameter estimates in logistic regression are about 1.6 to 1.8 times those in probit models.

Parameters in probit models can be interpreted in terms of effects on $E(Y^*)$ for the threshold latent variable model presented above. Since $Y^* = \alpha + \beta x + \epsilon$ where $\epsilon \sim N(0, 1)$ has cdf Φ , a 1-unit increase in x corresponds to a β increase in $E(Y^*)$. When ϵ is not in standardized form with $\sigma = 1$, a 1-unit increase in x corresponds to a β standard deviation increase in $E(Y^*)$. Alternatively, we can summarize effects on the probability scale, such as by comparing estimated probabilities at extreme values or quartiles of a predictor, with other predictors set at their means. (This was discussed for logistic models in Section 5.1.1.) Although probit model parameter estimates are on a different scale than logistic model parameter estimates, the probability summaries of effects are similar.

7.1.3 Probit Model Fitting

Let y_i be the number of successes out of n_i trials at setting x_i of possibly multiple explanatory variables, $i = 1, \dots, N$. Let x_{ij} denote the value of predictor, j for subject i . For the probit model $\Phi^{-1}[\pi(x_i)] = \sum_j \beta_j x_{ij}$ with $x_{i0} = 1$ and $\beta_0 = \alpha$, the log-likelihood function is

$$L(\boldsymbol{\beta}) = \log \left\{ \prod_{i=1}^N \left[\Phi\left(\sum_j \beta_j x_{ij}\right) \right]^{y_i} \left[1 - \Phi\left(\sum_j \beta_j x_{ij}\right) \right]^{n_i - y_i} \right\}.$$

Differentiation with respect to β_j leads to a special case of the likelihood [equations \(4.27\)](#) for binomial regression models,

$$\sum_i \frac{n_i [y_i - \Phi\left(\sum_j \beta_j x_{ij}\right)] x_{ij}}{\Phi\left(\sum_j \beta_j x_{ij}\right) [1 - \Phi\left(\sum_j \beta_j x_{ij}\right)]} \phi\left(\sum_j \beta_j x_{ij}\right) = 0,$$

with $\phi(\cdot)$ the standard normal pdf. When the link function is not the canonical one (which is the logit for binary data), there is no reduction of the data in sufficient statistics. Fisher (1935b), in an appendix to Bliss (1935) for the single predictor case, showed how to solve these equations using the algorithm now referred to as *Fisher scoring*. He also pointed out that cases with $y_i = 0$ or $y_i = n_i$ were not problematic for ML fitting, unlike weighted least squares using sample probits (or logits).

The estimated asymptotic covariance matrix of $\hat{\beta}$ has the GLM form [\(4.31\)](#)

$$\widehat{\text{cov}}(\hat{\beta}) = (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1}.$$

For probit models, $\hat{\mathbf{W}}$ is the diagonal matrix with elements

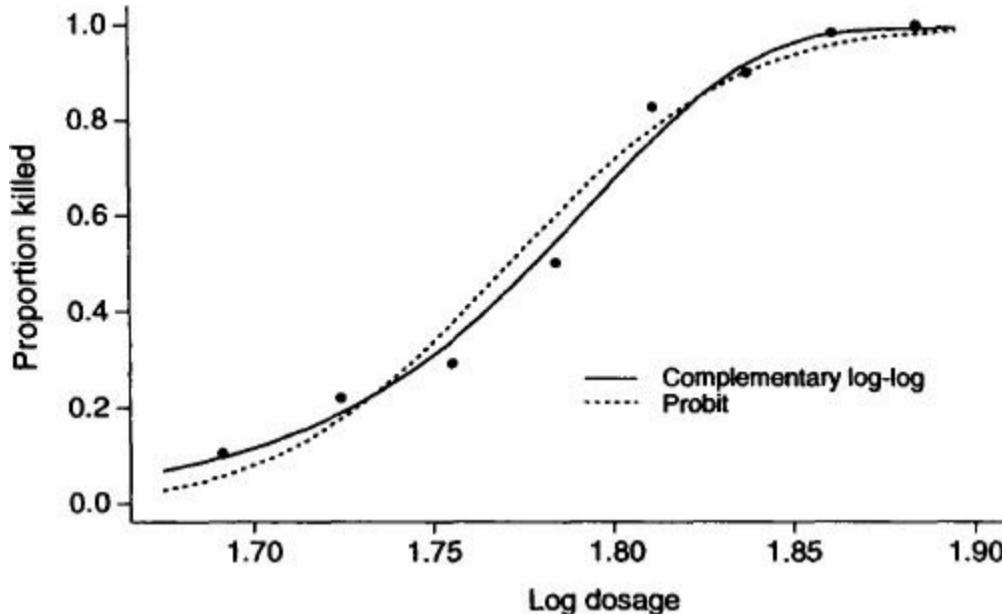
$$\hat{w}_i = n_i \left[\phi\left(\sum_j \hat{\beta}_j x_{ij}\right) \right]^2 / \left\{ \Phi\left(\sum_j \hat{\beta}_j x_{ij}\right) [1 - \Phi\left(\sum_j \hat{\beta}_j x_{ij}\right)] \right\}.$$

The Newton–Raphson algorithm yields the same ML estimates but slightly different standard errors. For the information matrix inverted to obtain the asymptotic covariance matrix, Newton–Raphson uses observed information, whereas Fisher scoring uses expected information. These differ for link functions other than the canonical link.

7.1.4 Example: Modeling Flour Beetle Mortality

[Table 7.1](#), from the Bliss (1935) article on probit modeling, reports the number of adult flour beetles killed after 5 hours of exposure to gaseous carbon disulfide at various concentrations. [Figure 7.1](#) plots (as dots) the proportion killed against the log concentration. The proportion jumps up at about $x = 1.8$, and it is close to 1 above there.

[Figure 7.1](#) Proportion of beetles killed versus log dosage, with fits of probit and complementary log-log models.



[Table 7.1](#) Beetles Killed After Exposure to Carbon Disulfide

Log Dose	Number of Beetles	Number Killed	Fitted Values		
			Comp. Log-Log	Probit	Logit
1.6907	59	6	5.6	3.4	3.5
1.7242	60	13	11.3	10.7	9.8
1.7552	62	18	21.0	23.5	22.5
1.7842	56	28	30.4	33.8	33.9
1.8113	63	52	47.8	49.6	50.1
1.8369	59	53	54.1	53.3	53.3
1.8610	62	61	61.1	59.7	59.2
1.8839	60	60	59.9	59.2	58.7

Source: Data reprinted with permission from Bliss (1935).

The ML fit of the probit model is

$$\Phi^{-1}[\hat{\pi}(x)] = -34.94 + 19.73x.$$

For this fit, $\hat{\pi}(x) = 0.50$ at $x = -\hat{\alpha}/\hat{\beta} = 34.94/19.73 = 1.77$. The fit corresponds to a normal tolerance distribution with $\mu = 1.77$ and $\sigma = 1/19.73 = 0.05$. The curve for $\hat{\pi}(x)$ is that of a $N(1.77, 0.05^2)$ cdf. As x increases from 1.6907 to 1.8839, the estimated probability of death increases from 0.057 to 0.987. For a 0.10-unit increase in x , such as from 1.70 to 1.80, we estimate that the conditional distribution of the latent variable y^* shifts up by $0.10(19.73) \approx 2$ standard deviations.

At dosage x_i with n_i beetles, $n_i\hat{\pi}(x_i)$ is the fitted count for death, $i = 1, \dots, 8$. [Table 7.1](#) reports the fitted values and [Figure 7.1](#) shows the fit. The table also shows fitted values for the linear logit model. These models fit similarly and rather poorly. The deviance G^2 goodness-of-fit statistic equals 11.23 for the logit model and 10.12 for the probit model, with $df = 6$. Bliss found an improved fit by combining a probit model for the lowest three concentrations with a separate one for the third through eighth concentration. We next consider an alternative model that gives a good fit to all eight concentrations at once.

7.1.5 Complementary Log–Log Link Models

The logit and probit links are symmetric about 0.50, in the sense that

$$\text{link}[\pi(x)] = -\text{link}[1 - \pi(x)].$$

To illustrate,

$$\begin{aligned}\text{logit}[\pi(x)] &= \log[\pi(x)/(1 - \pi(x))] \\ &= -\log[(1 - \pi(x))/\pi(x)] = -\text{logit}[1 - \pi(x)].\end{aligned}$$

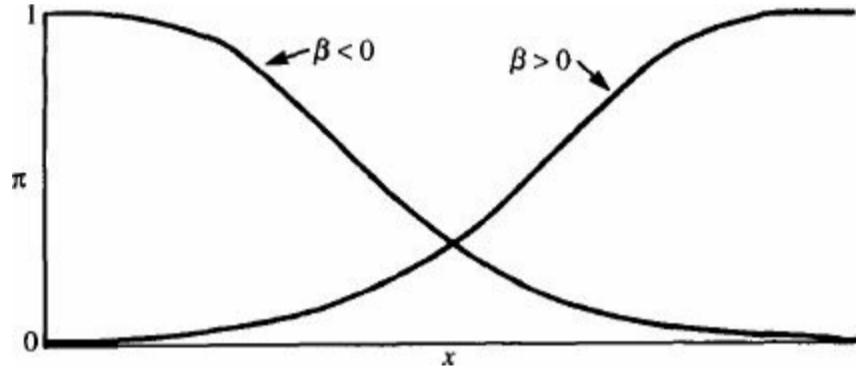
This means that the response curve for $\pi(x)$ has a symmetric appearance about the point where $\pi(x) = 0.50$, with $\pi(x)$ approaching 0 at the same rate that it approaches 1. Logistic models and probit models are inappropriate when this is badly violated.

The response curve

$$(7.3) \quad \pi(x) = 1 - \exp[-\exp(\alpha + \beta x)]$$

has the shape shown in [Figure 7.2](#). It is asymmetric, $\pi(x)$ approaching 0 fairly slowly but approaching 1 quite sharply. For this model,

[Figure 7.2](#) Binary regression model with complementary log–log link function.



$$\log[-\log(1 - \pi(x))] = \alpha + \beta x.$$

The link function for this GLM is called the *complementary log–log* link, since the log–log link applies to the complement of $\pi(x)$ (Yates 1955).

To interpret model [\(7.3\)](#), we note that at x_1 and x_2 ,

$$\log[-\log(1 - \pi(x_2))] - \log[-\log(1 - \pi(x_1))] = \beta(x_2 - x_1),$$

so that

$$\frac{\log[1 - \pi(x_2)]}{\log[1 - \pi(x_1)]} = \exp[\beta(x_2 - x_1)]$$

and

$$1 - \pi(x_2) = [1 - \pi(x_1)]^{\exp[\beta(x_2 - x_1)]}.$$

For $x^2 - x_1 = 1$, the complement probability at x^2 equals the complement probability at x_1 raised to the $\exp(\beta)$ power. As x increases, the curve is monotone increasing when $\beta > 0$.

A related model to [\(7.3\)](#) is

$$(7.4) \quad \pi(x) = \exp[-\exp(\alpha + \beta x)].$$

In GLM form it uses the *log–log* link function,

$$\log[-\log(\pi(x))] = \alpha + \beta x.$$

For it, $\pi(x)$ approaches 0 sharply but approaches 1 slowly. As x increases, the curve is monotone increasing when $\beta < 0$. When the log–log model holds for the probability of a success, the complementary log–log model holds for the probability of a failure.

Model [\(7.4\)](#) with log–log link is the special case of [\(7.2\)](#) with cdf of the *type I extreme value* (or *Gumbel*) distribution. The cdf equals

$$F(x) = \exp[-\exp[-(x - a)/b]]$$

for parameters $b > 0$ and $-\infty < a < \infty$. It has mode a , mean $a + 0.577b$, standard deviation $\pi b / \sqrt{6} = 1.283b$, and is highly skewed to the right. The term *extreme value* refers to this being the limit distribution of the maximum of a sequence of independent and identically distributed continuous random variables. Models with log–log links can be fitted using the Fisher scoring algorithm for GLMs.

7.1.6 Example: Beetle Mortality Revisited

For the flour beetle mortality data ([Table 7.1](#)), the complementary log–log model has ML estimates $\hat{\alpha} = -39.57$ and $\hat{\beta} = 22.04$. At dosage $x = 1.70$, the fitted probability of survival is $1 - \hat{\pi}(x) = \exp\{-\exp[-39.57 + 22.04(1.70)]\} = 0.885$, whereas at $x = 1.80$ it is 0.330 and at $x = 1.90$ it is 4×10^{-5} . The probability of survival at dosage $x + 0.10$ equals the probability at dosage x raised to the $\exp(22.04 \times 0.10) = 9.06$ power. For instance, $0.330 = (0.885)^{9.06}$. [Table 7.1](#) shows the fitted values and [Figure 7.1](#) shows the fit. They are close to the observed death counts ($G^2 = 3.45$, $df = 6$). The fit seems adequate.

The models with different link functions are not nested so cannot be compared with standard likelihood-ratio tests. The AIC values are 41.3 for the logit link, 40.2 for the probit model, 33.7 for the complementary log–log link, and 57.8 for the log–log link. These show a clear preference for the complementary log–log link. Aranda-Ordaz (1981) and Stukel (1988) proposed generalized link functions and also analyzed these data.

7.2 BAYESIAN INFERENCE FOR BINARY REGRESSION

Bayesian modeling of binary response variables provides an alternative to the frequentist modeling of Chapters 5 and 6. Our main focus here is on the probit and logistic regression models.

7.2.1 Prior Specifications for Binary Regression Models

Models can have many parameters, and a researcher may have more prior information about some of them than others. One simplistic approach takes the prior distribution for β to be constant over the multidimensional space of all possible parameter values. Then, the posterior distribution is a constant multiple of the likelihood function. That is, the posterior distribution is a scaling of the likelihood function so that it integrates out to 1. The mode of the posterior distribution is then the ML estimate. When the sample size is small or the data are unevenly distributed among the categories, the posterior distribution may be quite skewed rather than approximately normal. In such cases, the posterior mean can be quite different from the posterior mode and thus from the ML estimate.

Effect parameters in binary regression models can take value over the entire real fine. Then, such a flat prior distribution is *improper*, not integrating out to 1 over the space of possible parameter values.¹ A danger with improper prior distributions is that posterior distributions can also be improper for some models (Natarajan and McCulloch 1995). A Markov chain Monte Carlo (MCMC) algorithm for approximating the posterior distribution may fail to recognize that the posterior distribution is improper. Thus, it is safer to use a proper but relatively diffuse prior if you prefer the prior distribution to be flat relative to the likelihood function.

Considerable flexibility for a prior for β is provided by a multivariate normal density. A simple uninformative prior takes each mean to be 0, with a large standard deviation. If you use a common $N(0, \sigma^2)$ prior for each parameter, it is sensible to standardize the explanatory variables (e.g., with means of 0 and standard deviations of 1) so that the effects are comparable in interpretation. Otherwise, take the scale into account: For example, if x = time is reseated from years to months, the new parameter is $\frac{1}{12}$ th as large, so σ in the normal prior should be multiplied by $\sqrt{\frac{1}{12}}$ compared to when x is measured in years.

Using large σ in normal priors for β implies priors on the probability scale that are highly U-shaped, with about half the probability very close to 0 and half very close to 1. This seems intuitively to be rather informative, but in fact such priors have little influence, with the posterior looking much like the likelihood function. You could instead select σ so that the prior on an induced probability scale is close to uniform. With a single parameter, this is true except near the boundaries when $\sigma \approx 1.5$; using $\sigma = 1.69$ matches the normal to a uniform prior in the first two moments.

Some data analysts prefer a subjective Bayesian approach whereby prior distributions represent prior beliefs about β . For example, instead of using $\mu = 0$ and a very large σ for a $N(\mu, \sigma^2)$ prior distribution, you could take μ and σ such that $\mu \pm 3\sigma$ contains all values that have any plausibility for the parameter. If appropriate, you can also include correlation in the prior distribution between different parameters.

In practice, it is not obvious how to specify the hyperparameters for normal prior distributions for β . Data analysts think more easily in terms of plausible values for probabilities rather than for model parameters that pertain to a nonlinear function of the probabilities such as effects on the log odds. Alternatively, you can construct a prior distribution on the probability scale rather than a link function scale such as the logit, as we'll explain in Section 7.2.4. Or, many Bayesians prefer to use the Jeffreys prior, because of its invariance to the parameterization and other desirable properties. This prior density function relates to the information matrix, being proportional to $|\mathcal{J}|^{1/2}$. For binomial regression models, Ibrahim and Laud (1991) and Chen et al. (2008) showed that the Jeffreys prior is proper. With logit and probit link functions, this prior is symmetric and unimodal at 0. It and the corresponding posterior have thinner tails than any multivariate t distribution, and this holds also for the complementary log–log link.

The next example illustrates the potential impact of the choice of prior distribution. At this stage, we will not worry about the technical details of how to approximate the posterior distribution computationally, leaving this to Section 7.2.6.

7.2.2 Example: Risk Factors for Endometrial Cancer Grade

Heinze and Schemper (2002) described a study about endometrial cancer in which the purpose was to describe y = histology of 79 cases (0 = low grade for 30 patients, 1 = high grade for 49 patients) in terms of three supposed risk factors: x_1 = neovasculature (1 = present for 13 patients, 0 = absent for 66 patients), x_2 = pulsatility index of arteria uterina (ranging from 0 to 49), and x_3 = endometrium height (ranging from 0.27 to 3.61). [Table 7.2](#) shows some of the data. The complete data set is available at the text web site.

[Table 7.2](#) Part of Endometrial Cancer Data Set^a

HG	NV	PI	EH	HG	NV	PI	EH	HG	NV	PI	EH
0	0	13	1.64	0	0	16	2.26	0	0	8	3.14
...											
1	1	21	0.98	1	0	5	0.35	1	1	19	1.02

^a HG = histology grade, NV = neovasculature, PI = pulsatility index, EH = endometrium height.

Source: Data courtesy of Michael Schemper and Georg Heinze. Complete data ($n = 79$) at www.stat.ufl.edu/~aa/cda/cda.html.

For these data, we consider the main-effects model

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

using standardized versions of x_2 and x_3 . For all 13 patients having $x_1 = 1$, the outcome is $y = 1$. There is quasi-complete separation, and the ML estimate $\hat{\beta}_1 = \infty$. The 95% profile likelihood confidence interval for β_1 is $(1.28, \infty)$. Apparently, neovasculature is an important risk factor, so it is not sensible to drop it from the model because of its infinite estimate. With the Bayesian approach, the estimate of β_1 is finite.

In our Bayesian analyses, we use independent $N(\mu, \sigma^2)$ prior distributions for the model parameters. To reflect a lack of prior belief about the direction of the effects, we took each $\mu = 0.0$. Instead of the usual (0, 1) coding for the indicator variable x_1 , we let it take values -0.5 and 0.5 . The prior distribution is then symmetric in the sense that the logits for each group have the same prior variability as well as the same prior means, yet β_1 still has the usual interpretation of a conditional log odds ratio.

For these data, because the log likelihood is so flat in the β_1 dimension, posterior means for β_1 can be quite different for different prior distributions. To reflect a lack of information about the sizes of the effects, we first took the prior distributions to be quite diffuse, with $\sigma = 10$. The analysis can be implemented with Bayesian software such as WinBUGS or ordinary software that has a Bayes option (such as SAS PROC GENMOD, as shown at the text website), using an MCMC algorithm to approximate the posterior. [Table 7.3](#) shows posterior means, standard deviations, and 95% equal-tail posterior intervals, based on an MCMC process with 1,000,000 iterations. Chains were run with various starting values, and gave similar results. With such a long process, the Monte Carlo standard errors for the approximations to the Bayes estimates were negligible (about 0.001). [Table 7.3](#) also shows the ML results, for comparison.

[Table 7.3](#) Results of Fitting Models to Cancer Data Set of [Table 7.2](#)^a

Analysis	$\hat{\beta}_1$	SD	Interval	$\hat{\beta}_2$	SD	Interval	$\hat{\beta}_3$	SD	Interval
ML	∞	—	$(1.3, \infty)$	-0.42	0.44	$(-1.4, 0.4)$	-1.92	0.56	$(-3.2, -1.0)$
Bayes, $\sigma = 10$	8.93	4.78	$(2.1, 20.1)$	-0.47	0.45	$(-1.4, 0.4)$	-2.14	0.59	$(-3.4, -1.1)$
Bayes, $\sigma = 1$	1.65	0.69	$(0.3, 3.0)$	-0.22	0.33	$(-0.9, 0.4)$	-1.77	0.43	$(-2.7, -1.0)$

^a Interval is profile likelihood interval for ML and equal-tail posterior interval for Bayes.

Consider β_1 , for which the ML estimate is infinite. Based on the posterior mean, the estimated odds of the higher grade histology when neovasculature is present are $\exp(8.93) = 7555$ times the

estimated odds when neovascularization is absent. The 95% equal-tail posterior interval for β_1 is (2.11, 20.14). This provides the inference that $\beta_1 > 0$ and the effect seems to be large. The estimated size of the effect is imprecise, because of the flat log likelihood and the relatively disperse priors. Inferences about the model parameters were substantively the same as with the ML frequentist analysis.

For further comparison, we used more informative prior distributions. To reflect a stronger belief that the effects are not extremely strong, we took the prior standard deviations to be 1.0. Then nearly all the prior probability mass for the conditional odds ratio $\exp(\beta_1)$ falls between $\exp(-3.0) = 0.05$ and $\exp(3.0) = 20$. Results were quite different than with the ML frequentist analysis or the Bayesian analysis with $\sigma = 10$. The posterior mean for β_1 is now 1.65 instead of 8.93. Because $y = 1$ for all 13 patients having $x_1 = 1$, the frequentist approach tells us we cannot rule out any very large value for β_1 . By contrast, if we had strong prior beliefs that $|\beta_1| < 3$, then even with these sample results the Bayesian posterior inference has an upper bound of about 3 for β_1 .

Corresponding to the frequentist P -value for $H_a: \beta_1 > 0$, the Bayesian approach provides the posterior probability that $\beta_1 < 0$. This is 0.002 for the Bayesian analysis with $\sigma = 10$; that is, 0.0 is the 0.2 percentile of the posterior distribution. For this relatively flat prior distribution, this posterior tail probability is similar to the P -value of 0.001 for the one-sided frequentist likelihood-ratio test of $H_0: \beta = 0$ against $H_a: \beta > 0$, thus giving very strong evidence that $\beta > 0$. The posterior $P(\beta_1 < 0) = 0.007$ for the informative prior with $\sigma = 1.0$. With the more informative prior distribution centered at a lack of a treatment effect, this posterior probability provides a bit less evidence of a treatment effect.

Similar substantive results occur with corresponding probit models. For comparable results, prior σ values should be divided by about 1.6 to 1.8 to reflect the smaller variability on the probit link scale compared with the logit link.

7.2.3 Bayesian Logistic Regression for Retrospective Studies

The frequentist ML equivalence between prospective and retrospective logistic models has analogs for Bayesian methods. A key reference is Seaman and Richardson (2004). Their retrospective likelihood was combined with a Dirichlet prior distribution on exposure probabilities (for a discrete exposure variable or set of variables) in the control group. Their prospective likelihood was combined with an improper uniform prior distribution for the log odds that an individual with baseline exposure is diseased. They showed that the posterior distribution of log odds ratios is equivalent for the two approaches. Ghosh and Mukherjee (2010) surveyed Bayesian work on case-control studies. Topics of interest include measurement error, handling missingness, and flexibility for hierarchical structures.

7.2.4 Probability-Based Prior Specifications for Binary Regression Models

As an alternative to relying on normal prior distributions on a link function scale such as the logit or probit, you can construct a prior distribution on the probability scale. The chosen prior distribution then induces a corresponding prior distribution for the model parameters. Moreover, prior distributions for models with different link functions are then comparable, even though they are not identical on the scale of the parameters in the linear predictor.

This approach requires selecting prior distributions for at least as many probability values as there are parameters in the model. Suppose we choose M settings of explanatory variable values for placing prior distributions on the response probabilities. At setting s , denoted by $\mathbf{x}_{(s)}$ with $P(Y=1) = \gamma_{(s)}$ at that point, we select a prior distribution (such as a beta distribution) for $\gamma_{(s)}$. One way to indirectly determine the hyperparameters for a beta distribution is to guess two relevant values of the distribution, such as its mean and its standard deviation or a percentile such as the 95th percentile. Those values then determine the beta indices. The sum of the two beta indices for $\gamma_{(s)}$ corresponds to a particular number K_s of “prior observations” that the prior belief represents. That is, at setting s the beta prior density for $\gamma_{(s)}$ has parameters $K_s g_s$ and $K_s(1 - g_s)$, corresponding to mean g_s . Alternatively, we might specify the prior information using g_s and K_s . For example, we might indicate that our guess for $\gamma_{(s)}$ is $g_s = 0.30$ and that this prior information is relatively vague, being comparable to $K_s = 2$ prior observations, in which case the prior beta hyperparameters are 0.60 and 1.40. This type of prior is sometimes referred to as a *data augmentation prior* (Christensen et al. 2010, Sec. 8.4).

For simplicity, we treat these M prior distributions for the probabilities as independent. Then, the joint prior density function in terms of these M probabilities is

$$g(\gamma_{(1)}, \dots, \gamma_{(M)}) \propto \prod_s \gamma_{(s)}^{K_s g_s - 1} (1 - \gamma_{(s)})^{K_s(1-g_s) - 1}.$$

Suppose the link function for the model corresponds to the inverse of the cdf Φ , such as standard normal for the probit link and standard logistic for the logit link. Let ϕ denote the corresponding pdf. Then, in terms of the model parameters $\boldsymbol{\beta}^T = (\alpha, \beta_1, \beta_2, \dots)$, this prior density function corresponds to

$$g(\boldsymbol{\beta}) \propto \prod_s \{\Phi(\boldsymbol{\beta}^T \mathbf{x}_{(s)})\}^{K_s g_s - 1} [1 - \Phi(\boldsymbol{\beta}^T \mathbf{x}_{(s)})]^{K_s(1-g_s) - 1} \times \phi(\boldsymbol{\beta}^T \mathbf{x}_{(s)}).$$

7.2.5 Example: Modeling the Probability a Trauma Patient Survives

Bedrick et al. (1997) and Christensen et al. (2010, Chap. 8) illustrated this approach using data from 300 patients admitted to the University of New Mexico Trauma Center between 1991 and 1994. The response variable Y was whether the patient died ($1 = \text{yes}$, $0 = \text{no}$). The explanatory variables were $x_1 = \text{injury severity score}$ (taking values between 0 and 75), $x_2 = \text{trauma score based on a weighted average of several measurements such as systolic blood pressure and respiratory rate}$ (taking values from 0 for no vital signs to 7.84 for normal vital signs), $x_3 = \text{age}$, and $x_4 = \text{type of injury}$ ($1 = \text{penetrating, such as a gunshot or knife wound}$; $0 = \text{blunt, such as a car crash}$). The authors used a model that permitted the effect of type of injury to vary by age,

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5(x_3 x_4).$$

The data are available at the website for the Christensen et al. (2010) text.² Of the 225 patients with blunt injuries, 17 died, whereas of the 75 patients with penetrating injuries, 5 died.

To help in selecting prior distributions, the authors elicited percentile values for $P(Y = 1)$ from the trauma surgeon who provided the data, at six locations for settings of the explanatory variables. For example, the first location was $x_1 = 25$, $x_2 = 7.84$, $x_3 = 60$, and $x_4 = 0$, representing a person with normal vital signs who was not badly hurt. There, the chosen prior was beta(1.1, 8.5), which has mean 0.11 and standard deviation 0.10. By contrast, the third of the six locations, with $x_1 = 41$, $x_2 = 3.34$, $x_3 = 60$, and $x_4 = 1$, had a considerably more severe injury score and poorer trauma score. The beta(5.9, 1.7) prior chosen there has mean 0.78 and standard deviation 0.14. The six priors are highly informative (perhaps too much so), corresponding to adding 57.5 observations to the data when regarded as data augmentation priors.

These six prior distributions induce prior distributions for the logistic model parameters. With data augmentation priors such as these beta distributions, the posterior distribution has the shape of the likelihood function for the augmented data set. So, we can use standard frequentist software to find the posterior mode by finding the ML estimate for the augmented data.

For the chosen prior distributions. [Table 7.4](#) lists the posterior means and standard deviations of the logistic model parameters, as well as the corresponding ML estimates and their standard errors. The Bayes estimates differ somewhat in magnitude from the ML estimates, reflecting the informative prior distributions. However, substantive conclusions are similar. There is some indication that injury type has more of an effect at younger ages (e.g., the Bayes estimate of injury type is 0.9 at age 10 and about 0 at age 55), although the interaction is not statistically significant.

[Table 7.4](#) Bayesian and ML Fit of Logistic Regression Model for Trauma Data

Variable	Bayesian, Beta Priors		Bayesian, Normal Priors		Frequentist ML	
	Mean	Std. dev.	Mean	Std. dev.	Estimate	SE
Intercept	-1.79	1.10	-2.02	1.57	-2.061	1.526
Injury score	0.07	0.02	0.09	0.03	0.083	0.028
Trauma score	-0.60	0.14	-0.60	0.18	-0.553	0.171
Age	0.05	0.01	0.05	0.02	0.051	0.017
Injury type	1.10	1.06	1.44	1.41	1.338	1.334
Age × Injury type	-0.02	0.03	-0.01	0.03	-0.005	0.032

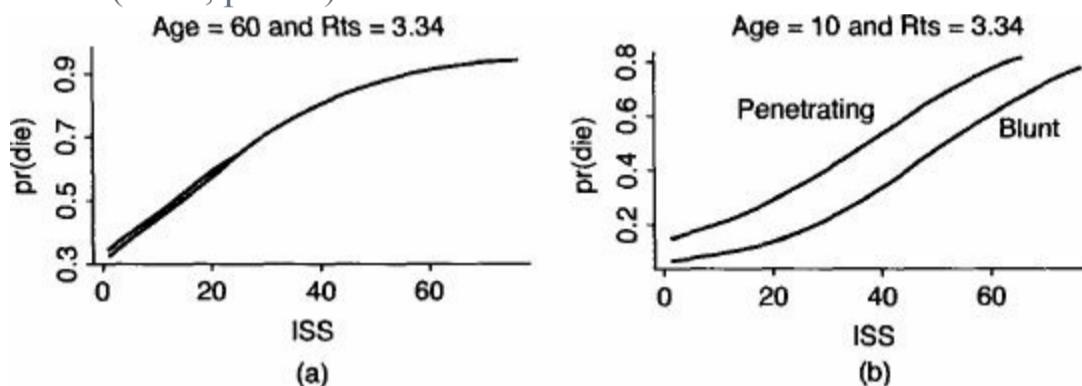
Source: Results with beta priors based on Table 2 in Bedrick et al. (1997).

This approach is appealing as a way of eliciting subjective priors that are interpretable on the probability scale. Some would find the resulting overall prior for the parameters as too highly informative. For comparison, [Table 7.4](#) also shows results from using relatively flat independent normal priors (each with $\sigma = 10$). These are more similar to those using ML. Alternatively, we could use the Bedrick et al. (1997) approach of setting priors on the probability scale to induce those for the logistic parameters, but select those probability priors to be less informative.

With the beta priors, Bedrick et al. (1997) reported that the posterior $P(\beta_1 < 0) < 0.01$. This corresponds to a frequentist P -value for testing $H_0: \beta_1 = 0$ against $H_1: \beta_1 > 0$. Not surprisingly, there is strong evidence that higher injury scores correspond to higher probabilities of non-survival.

At any particular setting of the explanatory variables, the posterior predictive value of $P(Y = 1)$ is found by integrating the logistic expression for this probability with respect to the posterior distribution for the model parameters. This gives a Bayesian posterior estimate of the probability of death at that setting. Bedrick et al. (1997) regarded this as more important than estimating the model parameters. Such estimates can be portrayed graphically as a way of describing effects. For example, [Figure 7.3](#) graphs the Bayes estimate of the probability of death as a function of the injury severity score (ISS), for each type of injury, for subjects with a trauma score (Rts) of 3.34 and ages of 10 and 60. This portrays how injury type has more of an effect at a younger age. Integrating the logistic curve by the posterior distribution yields a curve that does not have exactly the logistic formula, but has similar appearance. As in a frequentist analysis, it is also possible to provide interval estimates for $P(Y = 1)$.

Figure 7.3 Bayesian estimate of probability of dying as function of injury severity score and injury type, for subjects with trauma score = 3.34 and ages = 10 and 60. *Source:* Reprinted with permission from Christensen et al. (2010, p. 191).



Bedrick et al. (1997) studied the sensitivity of the results to deleting individual observations and to changes in the prior distribution. One way to summarize this is in terms of changes in the predictive value of $P(Y = 1)$ at the various explanatory settings for the observations. They also investigated link selection, by comparing models in terms of a Bayes factor (BF). The BF is formed as

$$BF = p(y|M_1)/p(y|M_2),$$

where $p(y|M)$ is the probability of the observed data under model M . For a given model M , $p(y|M)$ is obtained by integrating the likelihood function for that model with respect to the induced prior on β for that model, thus giving a *marginal likelihood*. For these data, this Bayes factor was about 1.0 when comparing models with logit and probit link, but about 21 when comparing each of these to the model with complementary log–log link. For example, for the chosen prior distributions, the probability of the observed data under the logistic model was 20.7 times the probability of the same data with the complementary log–log link.

7.2.6 Bayesian Fitting for Probit Models

We now summarize the basic ideas of Bayesian model fitting of binary regression models with normal priors, in the context of probit models. Albert and Chib (1993) showed that a simple analysis is possible in the probit case using Gibbs sampling based on the normal threshold latent variable model presented in Section 7.1.1. This model is simpler to handle than the logistic model, because results apply from Bayesian inference for ordinary normal linear regression models. Albert and Chib assumed a multivariate normal prior distribution for the regression parameters and independent normal latent variables. Then, the posterior distribution of the regression parameters, conditional on the observed data and the latent variables, is multivariate normal. Implementation of MCMC methods is relatively simple because the Monte Carlo sampling is from a normal distribution.

For subject i , a latent variable y_i^* is assumed to relate to the response y_i by $y_i = 1$ if $y_i^* > 0$ and $y_i = 0$ if $y_i^* \leq 0$. (We use the data in this section in ungrouped form, so that all $n_i = 1$.) We assume that y_i^* has a $N(\beta^T \mathbf{x}_i, 1)$ distribution, where $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^T$. Thus,

$$P(Y_i = 1) = P(Y_i^* > 0) = P(\beta^T \mathbf{x}_i + \epsilon > 0),$$

where ϵ is a $N(0, 1)$ random variable. But then for the standard normal cdf Φ , this corresponds to the probit model

$$\Phi^{-1}[P(Y_i = 1)] = \beta^T \mathbf{x}_i.$$

Albert and Chib noted that if $\{y_i^*\}$ were observed and a multivariate normal prior were chosen for β , then the posterior distribution for β results from ordinary normal linear model results. Given the binary response y_i , y_i^* is left or right truncated at 0, however. Thus, their distribution follows a truncated normal distribution, but it is still possible to use Gibbs sampling to simulate the exact posterior distribution.

The likelihood function can be constructed in terms of the model for the underlying latent observation, y_i^* . If y_i^* were observed, the contribution to the likelihood function would be $\phi(y_i^* - \beta^T \mathbf{x}_i)$, where $\phi(\cdot)$ is the standard normal pdf. Now, regarding y_i^* as unknown except for its sign, the contribution to the likelihood function is

$$[I(y_i^* > 0)^{y_i} I(y_i^* \leq 0)^{1-y_i}] \phi(y_i^* - \beta^T \mathbf{x}_i),$$

where I is the indicator function. For n independent observations, the likelihood function is proportional to the product of n such terms. Then, for prior density function $g(\beta)$, the joint posterior density of β and of $\{y_i^*\}$ given the data $\{y_i\}$ is proportional to

$$g(\beta) \prod_i [I(y_i^* > 0)^{y_i} I(y_i^* \leq 0)^{1-y_i}] \phi(y_i^* - \beta^T \mathbf{x}_i).$$

With the ML estimates as initial values, Albert and Chib used a Gibbs sampling scheme that successively samples from the density of $\mathbf{y}^* = (y_1^*, \dots, y_n^*)^T$ given β and of β given \mathbf{y}^* . With the conjugate normal prior, they noted that the posterior density of β given \mathbf{y}^* is normal. Specifically, suppose that the prior distribution of β is $N(\beta_0, \Sigma_0)$, and let \mathbf{X} be the matrix with i th row \mathbf{x}_i^T , so the latent variable model is $\mathbf{y}^* = \mathbf{X}\beta + \boldsymbol{\epsilon}$. Conditional on \mathbf{y}^* , the distribution of β is $N(\tilde{\beta}, \tilde{\Sigma})$ with

$$\tilde{\beta} = (\Sigma_0^{-1} + \mathbf{X}^T \mathbf{X})^{-1} (\Sigma_0^{-1} \beta_0 + \mathbf{X}^T \mathbf{y}^*), \quad \tilde{\Sigma} = (\Sigma_0^{-1} + \mathbf{X}^T \mathbf{X})^{-1}.$$

Conditional on β , the elements of \mathbf{y}^* are independent with density of y_i^* being $N(\beta^T \mathbf{x}_i, 1)$ truncated at the left by 0 if $y_i = 1$ and truncated at the right by 0 if $y_i = 0$. The model fitting yields posterior means for the Bayes estimates of parameters, and posterior standard deviations that describe the precisions of those estimates.

Albert and Chib also used a link function corresponding to the *cdf* of a t distribution, to investigate the sensitivity of fitted probabilities of response categories to the choice of link function. This approach provides the Cauchy link when $df = 1$ and the probit link as $df \rightarrow \infty$. It also provides close approximations to results for corresponding logistic models, because a t variate with $df = 8$ divided by 0.634 well approximates a standard logistic variate. They considered the t link case through a latent variable model using a scale mixture of normal distributions. They also considered a

hierarchical analysis that specifies priors for the hyperparameters of a normal prior distribution for β .

7.2.7 Bayesian Model Checking for Binary Regression

Model checking methods can investigate various aspects of the chosen model. Many of these methods parallel frequentist methods for model checking. For example, sensitivity analyses investigate how much posterior inferences change when alternative reasonable models are used. Case deletion diagnostics summarize the influence of individual observations. Bayes factors can be formed to compare different link functions in terms of posterior odds for a pair of models.

If the model is adequate, new data sets generated from the model should look like the observed data. Analogs of test statistics compare the observed data to predictive simulations based on the model. Analogs of P -values find the probability that replicated data are more extreme than the observed data.

Details of model checking methods are beyond the scope of this text. See Christensen et al. (2010, Sec. 8.3), Gelman et al. (2004, Chap. 6), and Spiegelhalter et al. (2002) and references therein. The Spiegelhalter et al. (2002) article proposed a complexity measure for the effective number of parameters in a model. They also proposed a *deviance information criterion* (DIC) for comparing models as a Bayesian analog of AIC. The DIC is based on adding double the effective number of parameters to a *mean posterior deviance* for checking fit. For a set of candidate models that seem to adequately explain the data, the model selected is the one that minimizes DIC.

7.3 CONDITIONAL LOGISTIC REGRESSION

ML estimators of logistic regression model parameters perform well when the sample size n is large compared with the number of parameters. When n is small or when the number of parameters grows as n does, improved inference results using *conditional maximum likelihood*.

The conditional likelihood approach eliminates nuisance parameters by conditioning on their sufficient statistics. The conditional likelihood refers to a distribution defined for potential samples that provide the same information about the nuisance parameters that occurs in the observed sample. We next introduce this approach, using it later in the text (Sections 16.5 and 16.6) for other small-sample inference in contingency tables. We'll also find it to be useful for the modeling of matched-pairs data in Section 11.2 and in more general contexts for clustered data in Section 13.1, for models in which the number of parameters grows as n does. In this setting it is an alternative to hierarchical Bayesian and frequentist random effects approaches that reduce the dimension of the parameter space by assuming a probability distribution for parameter sets that grow with the sample size.

7.3.1 Conditional Likelihood

We begin with a general exposition and then present special cases. For subject i , let y_i denote the binary response and let x_{ij} be the value of predictor j , $j = 1, \dots, p$. (For now, each y_i refers to a single trial, so $n_i = 1$.) The model is

$$(7.5) \quad P(Y_i = y_i) = \frac{\exp [y_i(\alpha + \sum_{j=1}^p \beta_j x_{ij})]}{1 + \exp (\alpha + \sum_{j=1}^p \beta_j x_{ij})}.$$

Substituting $y_i = 1$ gives the usual expression, such as (5.16). Here, we explicitly separate the intercept from the coefficients of the p predictors. For N independent observations,

$$(7.6) \quad P(Y_1 = y_1, \dots, Y_N = y_N) = \frac{\exp [(\sum_i y_i)\alpha + \sum_{j=1}^p (\sum_i y_i x_{ij})\beta_j]}{\prod_i [1 + \exp (\alpha + \sum_{j=1}^p \beta_j x_{ij})]}.$$

The sufficient statistic for α is $\sum_i y_i$, the total number of successes. The sufficient statistic for β_j is $\sum_i y_i x_{ij}$, $j = 1, \dots, p$.

Usually, some parameters refer to effects of primary interest. Others may be there to adjust for relevant effects, but their values are not of special interest. We can eliminate the latter parameters from the likelihood by conditioning on their sufficient statistics. We illustrate by eliminating α . (In Section 11.2.5 we will show that for models for matched case-control studies, there is a large number of intercept terms and they cause difficulties with inference about the primary parameters, so it is helpful to eliminate them.) The sufficient statistic for α is $\sum_i y_i$, so we condition on $\sum_i y_i$. Suppose that $\sum_i y_i = t$. Denote the conditional reference set of samples having the same value of $\sum_i y_i$ as observed by

$$S(t) = \left\{ (y_1^*, \dots, y_N^*): \sum_i y_i^* = t \right\}.$$

With $\{y_i\}$ such that $\sum_i y_i = t$, the conditional likelihood function equals

$$\begin{aligned} P(Y_1 = y_1, \dots, Y_N = y_N | \sum_i y_i = t) &= \frac{P(Y_1 = y_1, \dots, Y_N = y_N)}{\sum_{S(t)} P(Y_1 = y_1^*, \dots, Y_N = y_N^*)} \\ &= \frac{\exp [t\alpha + \sum_{j=1}^p (\sum_i y_i x_{ij})\beta_j] / \prod_i [1 + \exp (\alpha + \sum_{j=1}^p \beta_j x_{ij})]}{\sum_{S(t)} \exp [t\alpha + \sum_{j=1}^p (\sum_i y_i^* x_{ij})\beta_j] / \prod_i [1 + \exp (\alpha + \sum_{j=1}^p \beta_j x_{ij})]} \\ &= \frac{\exp [\sum_{j=1}^p (\sum_i y_i x_{ij})\beta_j]}{\sum_{S(t)} \exp [\sum_{j=1}^p (\sum_i y_i^* x_{ij})\beta_j]}. \end{aligned}$$

This does not depend on α .

Once we've obtained the conditional likelihood, we use it like an ordinary likelihood. For the parameters in it, their conditional ML estimates are the values maximizing it. Found using iterative methods, the estimators are asymptotically normal with covariance matrix equal to the negative inverse of the matrix of second partial derivatives of the conditional log likelihood. Likewise, we can construct large-sample Wald, likelihood-ratio and score tests using approximate chi-squared sampling distributions, and we can invert such tests to construct confidence intervals.

7.3.2 Small-Sample Inference for a Logistic Regression Parameter

As an alternative to large-sample methods, we can use the conditional distribution to perform “exact” inference, using permutation methods that consider the set of all data arrays that have the fixed values for the sufficient statistics we condition upon. With it, probabilities such as P -values occur exactly rather than as crude approximations (Cox 1970).

For instance, suppose that inference focuses on β_p in model (7.5). To eliminate other parameters, we condition on their sufficient statistics $T_j = \sum_i y_i x_{ij}$, $j = 0, \dots, p - 1$ (where $x_{i0} = 1$). With an argument like that just shown, we obtain the conditional distribution

$$P(Y_1 = y_1, \dots, Y_N = y_N | T_j = t_j, j = 0, \dots, p - 1) \\ = \frac{\exp\left[\left(\sum_i y_i x_{ip}\right) \beta_p\right]}{\sum_{S(t_0, \dots, t_{p-1})} \exp\left[\left(\sum_i y_i^* x_{ip}\right) \beta_p\right]} = \frac{\exp(t_p \beta_p)}{\sum_{S(t_0, \dots, t_{p-1})} \exp(t_p^* \beta_p)},$$

where

$$S(t_0, \dots, t_{p-1}) = \{(y_1^*, \dots, y_N^*): \sum_i y_i^* x_{ij} = t_j, j = 0, \dots, p - 1\}.$$

This depends only on β_p . Inference for β_p uses the conditional distribution of its sufficient statistic, $T_p = \sum_i y_i x_{ip}$, given the others. Let $c(t_0, \dots, t_{p-1}, t)$ denote the number of data vectors in $S(t_0, \dots, t_{p-1})$ for which $T_p = t$. The conditional distribution of T_p is

$$(7.7) \quad P(T_p = t | T_j = t_j, j = 0, \dots, p - 1) = \frac{c(t_0, \dots, t_{p-1}, t) \exp(t \beta_p)}{\sum_u c(t_0, \dots, t_{p-1}, u) \exp(u \beta_p)},$$

where the denominator summation refers to the possible values u of T_p .

For testing $H_0: \beta_p = 0$, the conditional distribution simplifies. For $H_a: \beta_p > 0$ and observed $T_p = t_{\text{obs}}$, the exact conditional P -value is

$$\sum_{t \geq t_{\text{obs}}} P(T_p = t | T_j = t_j, j = 0, \dots, p - 1) = \frac{\sum_{t \geq t_{\text{obs}}} c(t_0, \dots, t_{p-1}, t)}{\sum_u c(t_0, \dots, t_{p-1}, u)}.$$

This is the proportion of data configurations in the conditional set for which the sufficient statistic for β_p is at least as large as observed. Implementing this inference requires calculating $\{c(t_0, \dots, t_{p-1}, u)\}$. For all but the simplest problems, computations are intensive and require specialized software (e.g., LogXact of Cytel Software or PROC LOGISTIC in SAS).

7.3.3 Small-Sample Conditional Inference for 2×2 Contingency Tables

We illustrate first with a simple special case. Consider logistic regression with a single binary predictor x ,

$$(7.8) \text{logit}[P(Y_i = 1)] = \alpha + \beta x_i, \quad i = 1, \dots, N,$$

where $x_i = 1$ denotes row 1 and $x_i = 0$ denotes row 2. The model applies to a 2×2 table. The sufficient statistic $\sum_i y_i$ for α is the first column total. The sufficient statistic $T = \sum_i y_i x_i$ for β simplifies to the number of successes in the first row. Equivalently, the sufficient statistic for the model are the numbers of successes in the two rows. Let t and s denote these binomial variates. The row totals n_1 and n_2 are their indices.

To eliminate α , we condition on $\sum_i y_i = t + s$, the first column total. Since $N = n_1 + n_2$ is fixed, so then is the other column marginal total. Fisher (1935c) showed that fixing both sets of marginal totals yields *noncentral hypergeometric* probabilities for t that depend only on β ,

$$(7.9) \quad f(t|t+s; n_1, n_2, \beta) = \frac{\binom{n_1}{t} \binom{n_2}{s} e^{\beta s t}}{\sum_{u=m_-}^{m_+} \binom{n_1}{u} \binom{n_2}{s+t-u} e^{\beta u}}$$

for $m_- \leq t \leq m_+$ with $m_- = \max(0, n_{1+} + n_{+1} - n)$ and $m_+ = \min(n_{1+}, n_{+1})$. In that case the conditional distribution satisfies (7.7) with $c(t_0, t) = \binom{n_1}{t} \binom{N-n_1}{t_0-t}$ and with $t_0 = t + s$. The resulting exact conditional test that $\beta = 0$ is Fisher's exact test for 2×2 tables (Section 3.5.1).

7.3.4 Small-Sample Conditional Inference for Linear Logit Model

The linear logit model, $\text{logit}(\pi_i) = \alpha + \beta x_i$, applies to $I \times 2$ tables with ordered rows (Section 5.3.4). For it, the data $\{y_i\}$ are I independent $\{\text{bin}(n_i, \pi_i)\}$ counts, with fixed row totals $\{n_i\}$. Conditioning on $\sum y_i$ and hence the column totals yields a conditional likelihood free of α (Cox 1958). Exact inference about β uses its sufficient statistic, $T = \sum_i x_i y_i$. From (7.7) its distribution has the form

$$(7.10) \quad P(T = t | \sum_i y_i = t_0; \beta) = \frac{c(t_0, t) e^{\beta t}}{\sum_u c(t_0, u) e^{\beta u}}.$$

Here, $c(t_0, u)$ equals the sum of $\left[\prod_i \binom{n_i}{y_i} \right]$ for all tables with the given marginal totals that have $T = u$.

When $\beta = 0$, the cell counts have distribution that is a special case of a multivariate hypergeometric distribution, to be shown in (16.27). To test $H_0: \beta = 0$, ordering the tables with the given margins by T is equivalent to ordering them by the Cochran–Armitage statistic. Thus, this test for the linear logit model is an exact trend test.

In Section 5.3.6 we applied the Cochran–Armitage test to [Table 5.3](#) on maternal alcohol consumption and infant malformation. Even though $n = 32,574$, the table is highly unbalanced, with both very small and very large counts. It is safer to use small-sample methods. For the exact conditional trend test with the same row scores as used there, the one-sided P -value for $H_a: \beta > 0$ is 0.0168. The two-sided P -value is 0.0172, reflecting asymmetry of the conditional distribution, given the marginal counts. We obtained a two-sided P -value of 0.010 with the large-sample Cochran–Armitage test.

7.3.5 Small-Sample Tests of Conditional Independence in $2 \times 2 \times K$ Tables

For $2 \times 2 \times K$ tables $\{n_{ijk}\}$, the Cochran–Mantel–Haenszel test of XY conditional independence uses a standardization of $\sum_k n_{11k}$. For the logistic model

$$(7.11) \quad \text{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z, \quad i = 1, 2, \quad k = 1, \dots, K,$$

this is the sufficient statistic for β , the effect of X . To conduct a small-sample test of $\beta = 0$, we need to eliminate the other model parameters. Constructing the likelihood reveals that the sufficient statistics for $\{\beta_k^Z\}$ are the column marginal totals $[n_{+jk}]$ in each partial table. When X and Z are predictors, it is natural to treat the numbers of trials $\{n_{i+k}\}$ at each combination of XZ values as fixed. Thus, small-sample inference about β conditions on the row and column totals in each stratum.

Conditional on the strata margins, an exact test uses $T = \sum_k n_{11k}$. Hypergeometric probabilities occur in each partial table for the independent null distributions of $\{n_{11k}, k = 1, \dots, K\}$. The product of the K mass functions gives the null joint distribution³ of $\{n_{11k}\}$. This determines the null distribution of T . For $H_a: \beta > 0$, the P -value is the null probability that $T \geq t_{\text{obs}}$, for the fixed strata marginal totals. Mehta et al. (1985) presented a fast algorithm. The test simplifies to Fisher's exact test when $K = 1$.

7.3.6 Example: Promotion Discrimination

[Table 7.5](#) refers to U.S. government computer specialists of similar seniority considered for promotion. The table cross-classifies promotion decision by employee's race, considered for three separate months. We test conditional independence of promotion decision and race, or $H_0: \beta = 0$ in model [\(7.11\)](#). The table contains several small counts. The overall sample size is not small ($n = 74$), but one marginal count (collapsing over month of decision) equals zero, so we might be wary of using the CMH test.

Table 7.5 Promotion Decisions by Race and by Month

Race	July Promotions		August Promotions		September Promotions	
	Yes	No	Yes	No	Yes	No
Black	0	7	0	7	0	8
White	4	16	4	13	2	13

Source: J. Gastwirth, *Statistical Reasoning in Law and Public Policy*. San Diego, CA: Academic Press, 1988, p. 266.

For $H_a: \beta < 0$, the probability of promotion was lower for black employees than for white employees. Given the margins of the partial tables in [Table 7.5](#), n_{111} can range between 0 and 4, n_{112} can range between 0 and 4, and n_{113} can range between 0 and 2. The total $T = \sum_k n_{11k}$ can range between 0 and 10. The sample data are the most extreme possible result in each case. The observed $\sum_k n_{11k} = 0$, and the P -value is the null probability of this outcome. Software provides $P = 0.026$. A two-sided P -value, based on summing the probabilities of all tables no more likely than the observed table, equals 0.056.

7.3.7 Discreteness Complications of Using Exact Conditional Inference

Like Fisher's exact test, exact conditional inference for logistic regression is conservative because of discreteness. This is especially true when n is small or the data are unbalanced, with most observations falling in a single column or row. Using mid P -values in tests and related confidence intervals reduces conservativeness.

A particular difficulty occurs when no other set of $\{y_i^*\}$ values has the same value as the observed data for a sufficient statistic $\sum_i y_i x_{ij}$ on which we condition. In that case the conditional distribution of the sufficient statistic is degenerate. The P -value for the exact test then equals 1.0. This commonly happens when at least one explanatory variable x_j whose effect is conditioned out for the inference is continuous, with unequally spaced observed values.

Finally, a limitation of the conditional approach is requiring sufficient statistics for the nuisance parameters. Reduced sufficient statistics exist only with GLMs that use the canonical link. Thus, for instance, the conditional approach works for logistic models but not probit models.

7.4 SMOOTHING: KERNELS, PENALIZED LIKELIHOOD, GENERALIZED ADDITIVE MODELS

So far in this text we've performed rather severe smoothings of categorical data, by producing fitted values satisfying a particular model. In Sections 1.6 and 3.6 we found that Bayesian methods can perform a weaker type of smoothing than this, for example, by shrinking cell proportions in a contingency table toward a simple model without explicitly assuming that model. In Section 7.2, we employed Bayesian fitting of binary regression models, essentially smoothing the ML fit in the direction of a prior distribution. This section presents frequentist ways of smoothing categorical data, mainly in the context of analyzing a binary response variable.

7.4.1 How Much Smoothing? The Variance/Bias Trade-off

Smoothing methods, in a sense, have more of a nonparametric fashion, as they base analyses on a more general structure. There is then less potential for incorrect conclusions because of model misspecification. However, in some ways the demands are greater: We need to choose among a potentially infinite number of forms relating the response variable to the explanatory variables, the number of parameters is also then potentially much larger, and overfitting is a danger.

As we explained in Section 3.3.8, the comparison between completely model-based and other methods is at the heart of the fundamental statistical trade-off between variance and bias. Using a particular model has the disadvantage of increasing the potential bias (e.g., a true probability differing from the value corresponding to fitting the model to the population); but, it has the advantage that the parsimonious decrease in the parameter space has the effect of decreasing the variance in estimating characteristics of interest.

The methods presented in this section provide a compromise, typically starting with a model but smoothing results in some way to adjust for ways the model may fail. All smoothing methods require input from the methodologist to control the degree of smoothness imposed on the data in order to deal with the bias/variance trade-off, whether it be determined by a smoothing parameter in a frequentist approach or a prior distribution in a Bayesian approach.

7.4.2 Kernel Smoothing

Kernel estimation is a smoothing method that in its basic form is completely non-model-based. It is useful for any type of data, providing a sort of nonparametric way to estimate a probability density or mass function without assuming a parametric distribution. Like Bayesian methods, to estimate a mean (such as a cell probability) at a particular point, it smooths the data by using not only the data at that point (such as a sample proportion) but also the data at other points.

First, consider estimating joint cell probabilities $\boldsymbol{\pi}$ in a multiway contingency table by smoothing the sample cell proportions \mathbf{p} . Let \mathbf{K} denote a square matrix containing nonnegative elements. Kernel estimates of $\boldsymbol{\pi}$ have the simple form

$$(7.12) \quad \tilde{\boldsymbol{\pi}} = \mathbf{K}\mathbf{p}.$$

The column totals of \mathbf{K} are taken to be 1, which forces the sum of elements in $\tilde{\boldsymbol{\pi}}$ to be 1, like \mathbf{p} . Such kernels are usually constructed to yield probability estimates of form

$$\tilde{\pi}_i = (1 - \lambda)p_i + \lambda(\text{smoother}_i),$$

where λ is a constant that controls the degree of smoothing. Greater λ provides more smoothing. The structure used for the smoother, as imbedded in \mathbf{K} in expression (7.12), incorporates the other observations, its form depending on whether variables are binary, nominal, or ordinal. For ordinal data, for example, the smoothing gives more weight to nearby cells and works well when true probabilities in nearby cells are similar.

This method can also smooth binary response data in a regression context, for example, for constructing a graph to portray the form of dependence of y on a predictor. Copas (1983) presented a simple method of this sort for a single quantitative explanatory variable x . Let $\phi(\cdot)$ denote a symmetric unimodal *kernel function*. This is usually taken to be a bell-shaped pdf, such as the standard normal. At any value x , the kernel smoothed estimate of $P(Y=1|X=x)$ is

$$(7.13) \quad \tilde{\pi}(x) = \frac{\sum_i y_i \phi[(x - x_i)/\lambda]}{\sum_i \phi[(x - x_i)/\lambda]},$$

where $\lambda > 0$ is a smoothing parameter. At any point x , the estimate $\tilde{\pi}(x)$ is a weighted average of the $\{y_i\}$. For the simple function $\phi(u) = 1$ when $u = 0$ and $\phi(u) = 0$ otherwise, $\tilde{\pi}(x_k)$ simplifies to the sample proportion of successes at $x = x_k$. Then, there is no smoothing. When ϕ is proportional to the standard normal pdf, $\phi(u) = \exp[-u^2/2]$, we get behavior approaching this by letting $\lambda \rightarrow 0$.

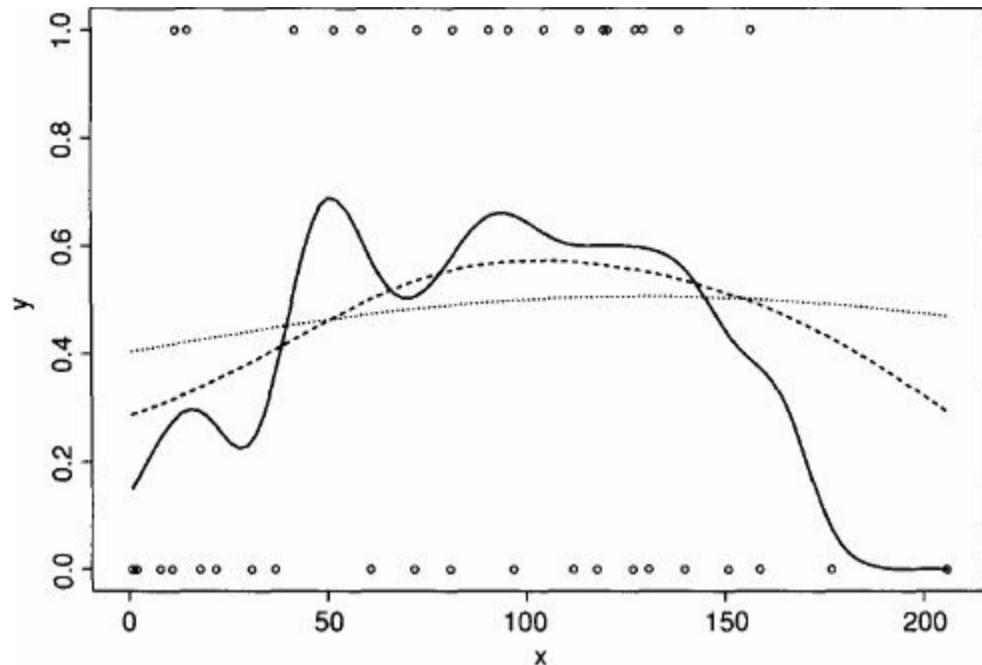
For very small λ , only points quite close to x have much influence. Then, using mainly very local data, there is little bias but high variance. By contrast, as λ increases, data points farther from x also can have a significant contribution to $\tilde{\pi}(x)$. As λ increases and more weight is given to points greatly distant, the smoother is more like the overall sample proportion, being more highly biased but with smaller variance. As λ grows unboundedly, the smooth function $\tilde{\pi}(x)$ converges to a horizontal line at the level of the overall sample proportion p of successes (Exercise 7.33).

For this kernel smoother, the choice of λ is more important to determining $\tilde{\pi}(x)$ than is the choice of ϕ . Copas recommended selecting λ by plotting the resulting function for several values of λ , varying around a value equal to 10 times the average spacing of the x values.

7.4.3 Example: Smoothing to Portray Probability of Kyphosis

Hastie and Tibshirani (1990, p. 282) described a study to determine risk factors for kyphosis, which is severe forward flexion of the spine following corrective spinal surgery. [Figure 7.4](#) shows this binary outcome y (1 = kyphosis present, 0 = absent) plotted against the age in months at the time of the operation. At the very low and very high levels of age, most observations have kyphosis absent.

[Figure 7.4](#) Kernel smoothing estimate of probability of kyphosis as function of age, using smoothing parameter $\lambda = 25$ (solid curve), 100 (dashed curve), 200 (dotted curve).



[Figure 7.4](#) also shows the result of kernel smoothing of the data using the smoother (7.13), with $\lambda = 25, 100$, and 200 . The value $\Lambda = 25$ is too low, and the figure is more irregular than the data justify. The higher values of Λ give evidence of nonmonotonicity in the relationship. In fact, adding a quadratic term to the standard logistic regression model provides an improved fit (Exercise 5.8).

7.4.4 Nearest Neighbors Smoothing

In more general contexts than binary regression, smoothers of the kernel type can base estimation at a point by using nearby points. A very simple such method is *nearest neighbors smoothing*. It is often used for classification, such as by predicting an observation for a subject based on a weighted average of observations for k subjects who have similar values on the explanatory variables.

Let s_{ij} be a measure of the similarity between subjects i and j , such as the Euclidean distance or Mahalanobis distance between values \mathbf{x}_i , and \mathbf{x}_j of explanatory variables for the two subjects, using standardized variables. For subject i , let $N(i)$ be the set of k subjects who are the nearest neighbors, having the k smallest values for s_{ij} among $j = 1, \dots, n$. Then, for a binary response, the probability $\pi_i = P(Y_i = 1)$ is estimated by

$$\hat{\pi}_i = \frac{\sum_{j \in N(i)} s_{ij} y_j}{\sum_{j \in N(i)} s_{ij}}.$$

Greater smoothing is produced by letting k be larger.

Sometimes the number of neighbors k to be used is fixed. Alternatively, cross-validation methods can be used to determine k . For each value of $k = 1, 2, \dots$, we could predict each observation using k neighbors, and then select the value of k for which the overall misclassification rate is smallest.

An advantage of this method is its simplicity, once we select a similarity measure to determine the nearest neighbors. However, the choice of this measure may not be obvious, especially when the number p of explanatory variables is large with possibly some subsets of them being highly correlated and some of them being qualitative. Also, the decision boundary between the \mathbf{x} values for classifying subjects into one category and the \mathbf{x} values for classifying subjects into another category can be highly irregular, especially when k is small. By contrast, the decision boundary is quite simple for standard binary regression models and some other methods, such as the linear discriminant method described in Section 15.1.

More complex smoothers further generalize this idea, for example, by basing the prediction at a point by a weighted regression using nearby points, such as described later in Section 7.4.9. Such methods have better statistical properties, such as usually lower bias than kernel smoothing. However, simple kernel smoothing is usually adequate for providing a sense of the main features of the true relationship.

7.4.5 Smoothing Using Penalized Likelihood Estimation

Kernel smoothing does not assume a probability distribution for Y or account for any dependence among the observations. Other methods do this, for example, by adjusting a likelihood function in such a way to induce smoothing. Consider an arbitrary model with generic parameter β and log-likelihood function $L(\beta)$. The *penalized likelihood* estimator of β maximizes

$$L^*(\beta) = L(\beta) - \lambda(\beta),$$

where $\lambda(\cdot)$ is a function that provides a roughness penalty. That is, $\lambda(\cdot)$ is such that $\lambda(\beta)$ decreases as elements of β are smoother in some sense, such as uniformly closer to 0.

First, consider smoothing a sparse contingency table, in which β are multinomial cell probabilities π . For two-way tables Simonoff (1983) suggested using a penalized likelihood approach with penalty

$$\lambda(\pi) = \lambda \sum_i \sum_j [\log(\pi_{ij}\pi_{i+1,j+1})(\pi_{i+1,j}\pi_{i,j+1})]^2$$

for the local odds ratios. This seems especially sensible with ordinal variables. It provides shrinkage toward the independence estimator, for which the local log odds ratios equal 0. To select the smoothing parameter λ , one approach minimizes an approximation for the mean squared error of the estimator.

In a more general modeling context, penalized likelihood applies with standard models such as logistic regression. For models incorporating standardized versions of explanatory variables, Lee and Silvapulle (1988) and le Cessie and van Houwelingen (1992) used a quadratic penalty term of form $\lambda(\beta) = \lambda \sum_j \beta_j^2$. This has the effect of shrinking estimates toward 0 and reducing prediction error.

Penalized likelihood methods are examples of *regularization methods*. These are ways of modifying ML methods to give sensible answers in situations that are unstable in some way, such as modeling using data sets containing very large numbers of variables. Regularization methods that penalize by a term that is quadratic in β , such as $\lambda(\beta) = \lambda \sum_j \beta_j^2$, are referred to as *L_2 -norm* methods. They are analogs of *ridge regression* for normal-response models. By contrast, *L_1 -norm* regularization uses penalty $\lambda(\beta) = \lambda \sum_j |\beta_j|$. Equivalently, it maximizes the likelihood subject to the constraint that $\sum_j |\beta_j| \leq K$ for some constant K . In ordinary regression, this penalty method is referred to as the *lasso* (“least absolute shrinkage and selection operator”). Another possible penalty, using *L_0 -norm*, takes $\lambda(\beta)$ to be proportional to the number of nonzero β_j . This approach has AIC and BIC as special cases. This sounds ideal, but optimization with this criterion is impractical with large numbers of variables; for example, the function minimized may not be concave. A compromise method, SCAD (“smoothly clipped absolute deviation”), starts at the origin $\beta = 0$ like a L_1 penalty and then gradually levels off (Fan and Lv 2010).

As in kernel smoothing, with penalized likelihood the degree of smoothing depends on the smoothing parameter λ , the choice of which reflects the bias/variance trade-off. Increasing λ results in greater shrinkage toward 0 in the estimates of $\{\beta_j\}$ and smaller variance but greater bias. Cross-validation criteria for selecting λ are based on fitting the model to part of the data and then checking goodness of fit in terms of predictions for the rest of the data. With k -fold cross-validation, this is done k times, each time leaving out the fraction $1/k$ of the data and predicting it using the model fit for the remaining data. The selected value of λ is the one for which the estimates have the lowest average prediction error, in some sense. That λ value is then used with the method applied to all the data.

Any particular norm for the penalty function has advantages and disadvantages. Use of a quadratic penalty is not a strategy for finding a parsimonious model, because all explanatory variables remain in the model. By contrast, with the lasso (L_1 penalty), when λ is large some β_j shrink to zero. With it,

It is informative to plot the estimates as a function of λ , to summarize how explanatory variables drop out as λ increases. For a factor predictor, the ordinary lasso solution may select individual indicators rather than entire factors, and the solution may depend on the coding scheme, so an alternative *grouped lasso* should be used. Disadvantages of the lasso approach compared with quadratic penalties are that it may overly penalize β_j that are truly large, and the $\{\hat{\beta}_j\}$ are not asymptotically normal and can be highly biased, making inference difficult. Some research on the lasso and related approaches has focused on adjusting the penalty function to make inference possible, to better determine which predictors truly have effects and to eliminate those that do not, and to penalize less severely when $|\beta_j|$ is large. For example, the SCAD approach puts little penalty on an effect that is estimated to be large but also has the effect of equating small coefficients to 0. An alternative penalty function has both L_1 and L_2 terms, each with its own penalty function. For binary data. Notes 7.7 and 7.8 cite several articles that have proposed and evaluated penalized likelihood methods.

Penalized likelihood estimators have connections with Bayesian smoothing methods. With prior density function proportional to $\exp[-\lambda(\beta)]$, the posterior density is proportional to the penalized likelihood function. Hence, the mode of the posterior distribution equals the penalized likelihood estimator.

7.4.6 Why Shrink Estimates Toward 0?

To methodologists who are used to estimators that are unbiased or approximately so, methods such as penalized likelihood that shrink $\hat{\beta}_j$ toward 0 can seem counterintuitive. Here is some intuition about why shrinkage may be effective. First, consider settings having a large number of explanatory variables for which most of them may have no effects or very minor effects, as in many genetics applications such as discussed in Section 7.5. Unless n is very large, by ordinary sampling variability ML estimates $\hat{\beta}_j$ will tend to be much larger in absolute value than the true values. This tendency is exacerbated when we consider only statistically significant values. Shrinkage such as occurs with penalized likelihood methods tends to move such estimates closer to the true values.

Second, variable selection methods such as the stepwise procedures discussed in Section 6.1.3 are highly discrete, in the sense that any particular variable either is or is not selected. Penalized likelihood is more continuous in nature, with some variables perhaps receiving little influence in the resulting prediction equation but not being completely eliminated. With the lasso method, a variable could be eliminated, but in a more objective way that is not dependent on which variables were previously eliminated.

7.4.7 Firth's Penalized Likelihood for Logistic Regression

Penalizing a likelihood need not necessarily increase bias. One version actually has been shown to reduce bias of ML estimators (Firth 1993a). For most models the ML estimator $\hat{\beta}$ has bias on the order of $1/n$, and Firth showed how to penalize the log likelihood such that this reduces to order $1/n^2$. For the canonical parameter of an exponential family model, the penalized log-likelihood function utilizes the determinant of the information matrix \mathcal{J} ,

$$L^*(\beta) = L(\beta) + \frac{1}{2} \log |\mathcal{J}|.$$

For application to logistic regression, Firth noted that when the model matrix is of full rank, $\log |\mathcal{J}|$ is strictly concave. Maximizing the penalized likelihood yields a maximum penalized likelihood estimate that always exists and is unique. This penalized likelihood then is proportional to the Bayesian posterior distribution resulting from using the Jeffreys prior. Thus, this penalized ML estimator equals the mode of the posterior distribution induced by the Jeffreys prior.

7.4.8 Example: Complete Separation but Finite Logistic Estimates

One situation in which Firth's penalized likelihood estimate is very helpful is when complete or quasi-complete separation occurs in the space of explanatory variables. Then, ordinary ML estimates of logistic regression parameters are infinite or do not exist (Section 6.5.1), but the penalized estimator is finite. Heinze and Schemper (2002) discussed this case.

We illustrate with the data from [Table 7.2](#) on risk factors for the histological grade of endometrial cancer, analyzed with Bayesian methods in Section 7.2.2. For the model

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

there is quasi-complete separation, and the ML estimate $\hat{\beta}_1 = \infty$. With standardized x_2 and x_3 , the other ML estimated effects are $\hat{\beta}_2 = -0.42$ ($SE = 0.44$) and $\hat{\beta}_3 = -1.92$ ($SE = 0.56$).

For comparison, the Firth penalized likelihood estimates are $\hat{\beta}_1 = 2.93$ ($SE = 1.55$), $\hat{\beta}_2 = -0.35$ ($SE = 0.40$), and $\hat{\beta}_3 = -1.72$ ($SE = 0.51$). The 95% profile penalized likelihood confidence interval for β_1 is $(0.61, 7.85)$, which shrinks the ordinary profile likelihood interval of $(1.28, \infty)$ considerably toward 0. Results for the other two estimates do not change as much.

The penalized likelihood estimates are posterior modes for the Bayesian approach using the Jeffreys prior. Compared with the posterior means for the normal priors reported in [Table 7.3](#), they fall between the results for normal priors with $\sigma = 1$ and with $\sigma = 10$. In this case, independent normal priors having $\sigma = 2$ provide a similar posterior mean for β_1 as Firth's penalized estimate.

7.4.9 Generalized Additive Models

The GLM generalizes the ordinary linear model to permit nonnormal distributions and modeling functions of the mean. The quasi-likelihood approach (Section 4.7) generalizes GLMs, specifying how the variance depends on the mean without assuming a particular distribution. Another generalization of the GLM replaces the linear predictor by additive smooth functions of the predictors. The GLM structure $g(\mu_i) = \sum_j \beta_j x_{ij}$ then generalizes to

$$g(\mu_i) = \sum_j s_j(x_{ij}),$$

where $s_j(\cdot)$ is an unspecified smooth function of predictor j . A useful smooth function is the *cubic spline*. It has separate cubic polynomials over sets of disjoint intervals, joined together smoothly at boundaries of those intervals. The boundary points, called *knots*, could be at evenly spaced points for each predictor or selected according to some criterion involving both smoothness and closeness of the spline to the data.

Like GLMs, this model specifies a link function g and a distribution for the random component. The resulting model is called a *generalized additive model*, symbolized by GAM (Hastie and Tibshirani 1990). The GLM is the special case in which each s_j is a linear function. Also possible is a mixture of explanatory terms of various types, with some s_j as smooth functions, others as linear functions as in GLMs, and others as indicator variables to include qualitative factors.

The details for fitting GAMs are beyond our scope. The *backfilling algorithm* employs a generalization of the Newton–Raphson method that utilizes local smoothing. The algorithm initializes $\{\hat{s}_j\}$; identically at 0. Then, at a particular iteration, it updates the estimate \hat{s}_j by a smoothing of the $\{y_i - \sum_{k \neq j} \hat{s}_k(x_{ik})\}$ that uses the other estimated smooth functions at that iteration, in turn for $j = 1, \dots, p$. The fitting procedure corresponds to subtracting from the log-likelihood function a penalty function that increases as the smooth function gets more wiggly.

The model fit assigns a deviance contribution and an approximate df value to each s_j in the additive predictor, enabling inference about those terms. For instance, a smooth function having $df = 4$ is similar in overall complexity to a third-degree polynomial, which has four parameters. Choosing a df value or a value for a smoothing parameter determines how smooth the resulting GAM fit looks. As in GLMs, we can compare deviances for nested models to test whether a model gives a significantly better fit than a simpler model. A disadvantage compared with GLMs is the loss of interpretability for describing effects of an explanatory variable that has a smooth term in the model.

It is usually sensible to try various degrees of smoothing to find one that smooths the data sufficiently so that the trend is not too irregular but does not smooth so much that it suppresses interesting patterns. The smoothing may suggest that a linear model is adequate with a particular link function or it may suggest ways to improve on linearity. Some software packages that do not have GAMs can smooth the data by employing a type of regression that gives greater weight to nearby observations in predicting the value at a given point; such *locally weighted least-squares regression* is often referred to as *lowess*. We prefer GAMs because they recognize explicitly the form of the response variable. For instance, with a binary response, lowess can give predicted values below 0 or above 1 at some predictor settings. This cannot happen with a GAM that assumes a binomial random component.

Even if you plan to use GLMs, a GAM is helpful for exploratory analysis. For instance, for continuous x with continuous responses, scatter diagrams provide visual information about the dependence of y on x . For binary responses, such diagrams are not very informative. Plotting the fitted smooth function for a predictor may reveal a general trend without assuming a particular functional relationship.

7.4.10 Example: GAMs for Horseshoe Crab Mating Data

For the horseshoe crab data introduced in Section 4.3.2, [Figure 4.4](#) showed the trend relating a female crab's number of male satellites to the width of her carapace shell. This smooth curve is the fit of a generalized additive model, assuming a Poisson distribution and using the log link.

In Section 5.1.3 we used logistic regression to model the probability that a female crab has at least one male satellite. For crab i , $y_i = 1$ if she has at least one satellite and $y_i = 0$ otherwise. [Figure 5.2](#) plotted these data against the crab's carapace width. That figure also showed a curve based on smoothing the data using a GAM, assuming a binomial response and logit link. This curve shows a roughly increasing trend and is more informative than viewing the binary data alone.

7.4.11 Advantages/Disadvantages of Various Smoothing Methods

Compared with simple kernel smoothing, penalized likelihood methods and the GAM have the advantage that the ordinary GLM is a special case. They also have a direct inferential aspect, as they mimic GLMs in assuming a binomial distribution for a binary response and having a df value associated with each explanatory effect.

Compared with ordinary frequentist inference, all these methods have the extra aspect of choosing the degree of smoothing. With the Bayesian approach, this is handled with the choice of prior distribution. An advantage of the Bayesian approach is that its entire formulation has a stronger theoretical basis, whereas the other smoothing methods have somewhat of an ad hoc nature in their adaptive choice of a smoothing parameter. But, of course, some methodologists are uncomfortable with the Bayesian paradigm because of the need to impose the prior; they may instead prefer such frequentist methods or empirical Bayes methods that let the data determine the degree of smoothing.

7.5 ISSUES IN ANALYZING HIGH-DIMENSIONAL CATEGORICAL DATA

In an increasing variety of recent applications, data sets differ from traditional ones in having a very large number p of variables, sometimes even more than the number n of observations. In genomics, such applications include classifying tumors by using microarray gene expression or proteomics data or associating protein concentrations with expression of genes or predicting a clinical prognosis by using gene expression data. Variable selection is especially important in such applications, as typically most effects are expected to be null. Other applications having large p include biomedical imaging, functional magnetic resonance imaging, tomography, signal processing, image analysis, market basket data, and portfolio allocation in finance.

Traditional modeling can deal effectively with the sorts of examples shown in this book, such as modeling a response for a few drugs and several centers in a clinical trial, but it can be overwhelmed when it needs to address differential expression (i.e., change between two or more conditions) in many thousands of genes or brain activity in many thousands of locations. Special software packages⁴ are needed to organize and analyze the data. Methods presented in this chapter, such as regularization methods employing penalized likelihood, are increasingly used with high-dimensional data.

In this section we'll discuss the analysis of high-dimensional categorical data. It is impossible to do justice to the exploding literature in this area in a single section. We'll focus on a particular difficult issue that arises in many such studies—selecting explanatory variables for an analysis out of a very large set, and making adjustments for multiplicity. We'll then describe some applications in which novel approaches have been proposed for high-dimensional analyses.

7.5.1 Issues in Selecting Explanatory Variables

In modeling with a very large number of explanatory variables, reducing their number can ease interpretability and decrease prediction errors by removing variables that have little if any relevance. For example, in disease classification, of a large number of genes relatively few may be responsible for the disease. Most effects are null or essentially null. This is reflected by histograms of P -values for testing those effects, which often have appearance quite similar to a uniform density function. In addition, with large p , ordinary ML fitting may not even be possible. For a binary response, complete separation often occurs once the number of predictors exceeds a particular point. Even when finite estimates exist, they may be very imprecise because of ill-conditioning of the covariance matrix. Moreover, choosing a model that contains a large number of predictors runs the risk of overfitting the data. Then, future predictions will tend to be poorer than those obtained with a simpler model.

In regression modeling, variable selection algorithms such as forward selection and backward elimination are popular. However, such methods have potential pitfalls, especially when p is large. In particular, for the set of predictors having no true effect, the maximum sample correlation with the response can be quite large. Also, there can be spurious collinearity among the predictors or spurious correlation between an important predictor and a set of unimportant predictors, because of the dimensionality.⁵ Other criteria exist for identifying an optimal subset of explanatory variables, such as minimizing prediction error or (with AIC) minimizing divergence of the fitted model from reality. With large p , though, it is not feasible to check all possible subsets of predictors. Recently, various methods have been proposed to deal with the subset selection issue for large p . Roughly, the methods fall into two types.

One approach uses alternatives to ML estimation such as various penalized likelihood methods mentioned in Section 7.4.5. These include regularization using L_q -norm for some q between 0 and 2 and compromise norms. Zhu and Hastie (2004) applied this to logistic regression with $q = 2$ for microarray cancer diagnosis. Besides providing shrinkage of parameter estimates, some of those methods (L_q -norm with $0 \leq q \leq 1$) can also help with variable selection. With the lasso ($q = 1$), many of the explanatory variables receive zero weight in the prediction equation. The number of such variables included depends on the smoothing parameter. However, the lasso has a tendency to include many false positive variables when p is large (Fan and Lv 2010) and to exclude important suppressor variables (Magidson 2010). Note 7.8 cites several articles that have investigated penalized likelihood methods for variable selection.

A second approach attempts to identify the relevant effects using standard significance tests but with some adjustment for multiplicity. This can reduce dramatically the dimensionality of the data by eliminating the large number of predictors for which there is not strong evidence of an effect. This approach is especially useful in applications in which a small portion of the effects considered truly exist. We next discuss such multiplicity adjustments, including one (the *false discovery rate*) that has received substantial attention in recent years for large p applications.

7.5.2 Adjusting for Multiplicity: The Bonferroni Method

[Table 7.6](#) summarizes results of $g = n_{11} + n_{12} + n_{21} + n_{22}$ significance tests, of which there are n_{12} incorrect rejections of H_0 (i.e., type I errors) and n_{21} incorrect nonrejections of H_0 (type II errors). Testing each hypothesis at level α^* ensures that $E(n_{12}/g) \leq \alpha^*$. In practice, we observe the numbers $n_{+1} = (n_{11} + n_{21})$ of nonrejections and $n_{+2} = (n_{12} + n_{22})$ of rejections, but the actual cell counts are unknown.

Table 7.6 Contingency Table Summarizing Multiple Significance Tests. Type I and Type II Errors Have Frequencies n_{12} and n_{21} .

		Decision	
		Do Not Reject H_0	Reject H_0
Condition of H_0	H_0 true	n_{11}	n_{12}
	H_0 false	n_{21}	n_{22}

A substantial literature exists about ways of controlling error rates when conducting a large number of statistical inferences. In a testing format, in making multiple comparisons of groups on some response variable, the “familywise” error rate approach controls $P(n_{12} > 0)$, the probability of making at least one type I error. In a confidence interval format, this corresponds to having the confidence coefficient apply to the entire set of intervals formed rather than to each individual one.

As Section 3.1.8 explained, a simple multipurpose way to establish control over a family of inferences is the Bonferroni method. This method ensures a familywise error bound of $\alpha = P(n_{12} > 0)$ by setting $\alpha^* = \alpha/g$ for each inference. However, the method is conservative, having actual error rate bounded above by the nominal level α .

When g is enormous, such as in detecting differential expression in thousands of genes, the Bonferroni approach is *too* conservative because α/g is so tiny. This makes it difficult to establish significance in any one test and to discover any effects that truly are there. But, in the absence of such an adjustment, there is the danger that most significant results found will be type I errors, because of the relatively small number of true effects. Dudoit et al. (2003) described many alternatives to the Bonferroni method, perhaps the most popular of which we present next.

7.5.3 Adjusting for Multiplicity: The False Discovery Rate

In [Table 7.6](#), consider the ratio n_{12}/n_{+2} , which is the proportion of the rejected null hypotheses that are erroneously rejected. Then, $FDR = E(n_{12}/n_{+2})$ where we set $n_{12}/n_{+2} = 0$ when $n_{+2} = 0$, is called the *false discovery rate* (Benjamini and Hochberg 1995).

Suppose all null hypotheses are true. If $n_{12} = 0$, then $n_{12}/n_{+2} = 0$. whereas if $n_{12} > 0$, then $n_{12}/n_{+2} = 1$, so that $FDR = P(n_{12} > 0)$, the same as the family wise error rate. When some null hypotheses are false, then FDR is less than the familywise error rate. So, if a procedure controls the FDR only, it can be less stringent and therefore less conservative. It is then more powerful, tending to yield more rejections of false H_0 's, more so as g increases and as the number of false hypotheses increases.

There are several variations on FDR and many types of FDR algorithms. For FDR as just defined, Benjamini and Hochberg (1995) suggested a simple algorithm for ensuring $FDR \leq \alpha$ for a desired α . It applies with g independent tests. Let $P_{(1)} \leq P_{(2)} \leq \dots \leq P_{(g)}$ denote the ordered P -values. Then, we reject the corresponding hypotheses $(1), \dots, (j^*)$, where j^* is the maximum j for which $P^{(j)} \leq j\alpha/g$. The most significant test compares $P_{(1)}$ to α/g and has the same decision as in the ordinary Bonferroni method, but then the other tests have less conservative requirements. The actual FDR for this method is bounded above by $(n_{1+}/g)\alpha$, which is α when the null is always true.

Benjamini and Hochberg illustrated their method for a study about myocardial infarction. For the 15 hypotheses tested, the ordered P -values were

0.0001, 0.0004, 0.0019, 0.0095, 0.020, 0.028, 0.030,

0.034, 0.046, 0.32, 0.43, 0.57, 0.65, 0.76, 1.00.

With $\alpha = 0.05$, these are compared with $j(0.05)/15$, starting with $j = 15$. The maximum j for which $P_{(j)} \leq j(0.0033)$ is $j = 4$, for which $P_{(4)} = 0.0095 < 4(0.0033)$. So, the hypotheses with four smallest P -values were rejected. By contrast, the Bonferroni approach with familywise error rate 0.05 compares each P -value to $0.05/15 = 0.0033$ and rejects only three of these hypotheses.

Benjamini and Yekutieli (2001) showed that this method works even when the tests are positively dependent in a certain sense. They suggested an adjusted method for general dependence structure, but it is more conservative. An alternative approach fixes a threshold for each test statistic value or P -value and then estimates the FDR for the set of tests (Lin et al. 2010). The FDR method also applies when the tests are discrete, under which the null distributions of P -values are not uniform but instead are stochastically greater than uniform. For discrete data, Gilbert (2005) improved the method by combining it with an adjustment for discreteness that Tarone (1990) had suggested for the Bonferroni method.

Because of its lessened conservatism and improved power compared with familywise methods such as Bonferroni, controlling FDR is a sensible strategy to employ in exploratory research involving large-scale testing. There is still then a place for traditional familywise multiple comparison methods in follow-up validation studies involving the smaller numbers of effects found to be significant in the exploratory studies. Dudoit et al. (2003) surveyed these issues, in the context of microarray experiments.

7.5.4 Other Variable Selection Methods with High-Dimensional Data

Fan and Lv (2010) surveyed many ways of dealing with large p by reducing the number of explanatory variables. Most of them incorporate at least one of the subset selection approaches discussed above—stepwise algorithms, regularization such as penalized likelihood methods, and adjustments for multiplicity.

One approach conducts a large-scale screening to eliminate unimportant variables and then a moderate-scale screening to select from them the important variables. For a quantitative predictor, the large-scale screening could use a two-sample t test to compare the mean responses for the two groups that are the categories for y . An alternative method uses a stepwise algorithm, such as forward selection. To reduce the problems mentioned in Section 7.5.1 for applying forward selection. Park and Hastie (2008) recommended implementing it with a penalized likelihood function using a quadratic penalty term.

An alternative variable reduction method replaces the set of explanatory variables by a much smaller set of artificial variables and then applies variable selection methods to them. With principal component regression, each artificial variable is a linear combination of the original variables, designed to explain as much variance as possible. Magidson (2010) proposed a related *correlated component regression* that bases the first component on an average of effects in single-predictor models, the second component on an average of effects in two-predictor models that use the first component as one of the two predictors and one of the explanatory variables as the second one, and so forth. With such methods, there is no guarantee that the new components will be predictive of the response variable, and their effects are not as interpretable, especially when p is very large. However, in screening out many of the explanatory variables before the components are formed, Magidson cautioned against screening out suppressor variables that may reveal their relevance only when some other variable is already in the model.

7.5.5 Examples: High-Dimensional Applications in Genomics

Fan et al. (2010) described several computational biology topics for which specialized high-dimensional analyses have recently been proposed. These include human genetics and disease mapping, discrete sequence motif discovery, protein sequence alignment, population genetics, evolutionary models, and finite mixture clustering for microarray data. The amount of molecular data is enormous, such as billions of base pairs of DNA sequence data in the GenBank, and much of the data is of categorical form. A main goal of many studies is to discover genetic variations that underlie whether or not a certain disease is present. Methods for categorical data analysis can determine whether an association exists between a person's genetic marker genotype and his/her disease status.

To discover genetic markers that may be associated with a disease, case-control structure is often used with cases and noncases of the disease (Li and Conti 2009, Umbach and Weinberg 1997). The most abundant variations in the human genome are the single-nucleotide polymorphisms (SNPs), and an association analysis can study a SNP's genotype frequency in a group of diseased patients and a group of controls. For a single SNP, the analysis might refer to a 2×3 table that cross-classifies (case, control) by the SNP possible pairs of alleles (AA, AB, BB) for an individual's genotype. But many studies for detecting genetic signals use hundreds of thousands of SNP markers genotyped for thousands of subjects. The effects detected are usually quite weak, with relative risks between about 1.1 and 1.5. Pathway-based approaches attempt to build power by examining whether test statistics for a group of related genes have consistent yet moderate deviation from chance. Yet, genetic risk prediction can be challenging even when combining information from various studies.⁶

A further complication in attempting to find SNP markers whose genotype frequencies are significantly different between cases and controls is that interactions among them may affect the disease risk (Zhang and Liu 2007, Zhang et al. 2011). Allowing interactions is also crucial in the building of models for risk prediction. Some markers may have negligible effects by themselves but considerable effect in combination with other markers. In such applications, the number of genotyped markers can be much larger than the number of subjects, and the potential number of possible interaction combinations can be astronomical while there may be one or relatively few of them that are associated with the disease.

Dudoit et al. (2003) noted that the biological question of differential expression is a multiple hypothesis testing problem: simultaneously testing for each of possibly thousands of genes of the null hypothesis of no association between the expression levels and the responses or covariates. In genetic association studies, the null hypothesis is true or close to being true in a vast majority of cases. So, there is the multiplicity danger that most significant results found will be type I errors. Because of this, often extremely stringent sizes are used for significance, such as 5×10^{-8} instead of the usual 0.05. Multiplicity adjustments such as *FDR* are especially relevant, as is then replicating the finding of significance in an independent sample.

Yet another complication is that complex dependencies may exist among the tests applied to the many parameters for a particular data set, which makes the complete null distributions of test statistics and subsequent *P*-values unclear. Permutation methods can then be useful. An empirical distribution of *P*-values can be generated by repeating the tests with appropriate permutations of the data, such as by randomly permuting case-control labels, to induce a null distribution of the *P*-values while preserving the correlation structure. For any particular *P*-value, the proportion of the permutations that give a *P*-value smaller than the observed one provides an adjusted *P*-value (Dudoit et al. 2003).

Zhang et al. (2011) argued that for genetic association studies, ordinary penalized likelihood and stepwise selection methods (e.g., that identify the 10% most effective markers and then search for interactions among them) are ineffective. They summarized a Bayesian partitioning model based on a multinomial likelihood with Dirichlet prior. It is designed to partition SNPs into ones that are

unrelated to the disease, ones that are marginally associated with the disease, and ones that are jointly associated (in an interaction) with the disease. The output is a posterior probability for each SNP of belonging to each of these three groups. Zhang et al. (2011) gave an example in which an important interaction was detected between two SNPs for Crohn’s disease, with data containing gentotypes at 1182 SNPs for about 2000 cases and 3000 controls. As would be expected, posterior probabilities depend strongly on prior probabilities, but the order of posterior probabilities for different SNPs was little affected.

By contrast, Park and Hastie (2008) compared penalized likelihood incorporating a quadratic penalty function to other methods and found that it performs well in identifying relevant interactions in gene–environment interaction models. They found that it overcomes collinearity among predictors and can handle applications where the number of predictors is large relative to the sample size.

In some applications, analyses have combined several methods we’ve presented in this chapter. In modeling binary prostate cancer status, Liu et al. (2008) used a semiparametric logistic model with a linear effect of age but a nonparametric function of five genes within the cell growth pathway, maximizing a penalized binomial likelihood. For the nonparametric function, rather than using a smoothing spline such as in GAMs, the authors used a positive definite kernel function. The nonparametric approach for the part of the model involving genetic effects reflects the complex way that genes may interact with each other and relate to the response. The likelihood penalty and the kernel function both incorporate smoothing parameters. In a novel approach, rather than estimate these smoothing parameters by cross-validation or by trial-and-error inspection, Liu et al. (2008) treated them like variance components in a random effects model.

7.5.6 Example: Motif Discovery for Protein Sequences

Here is an example that illustrates the severity of the challenge in many genetics applications. Liu et al. (1995) proposed a multinomial mixture model for motif discovery for protein sequences, using probability vectors for particular motif sites mixed with probabilities that apply when an observation does not belong to any motif site. The data take the form of very long sequences of the four nucleotides which make up the DNA, commonly abbreviated by A, C, G, and T. Understanding many biological processes involves identifying relatively short patterns of these embedded in long strings. It is beyond our scope to explain the details, but the challenging aspect of it is shown by the general setting in which Liu et al. (1995) described the problem: Consider L coins, of which $L - J$ all have the same probability π of head, and the remaining J have probabilities of heads different from π and different from one another. In each of K independent trials, the L coins are flipped and laid out in a row such that J special coins are in a contiguous block in the same order in each trial, but with unknown location that can vary from trial to trial. The challenge is to estimate all the probabilities and identify the locations of the special coins in each trial.

More generally, there may be multiple possible outcomes for each coin (such as A, C, G, and T), there could be multiple blocks of the special coins in each trial, the special coins might not occur in a contiguous block, and it may be of interest to test the hypothesis that there are special coins with different probabilities from the common probability of the other coins. Liu et al. (1995) and Jensen et al. (2004) provided Bayesian solutions, modeling the sequence patterns as a product multinomial using Dirichlet priors and treating the motif finding problem as a missing data problem (because the motif locations are not observed). The motif sequence frequencies can be well estimated with a large number of fragments, but the number of possibilities grows exponentially with the number of fragments because of the unobserved locations.

7.5.7 Example: The Netflix Prize

Netflix is a company that distributes movies to subscribers through the mail with DVDs and by Internet streaming. For each movie viewed, a subscriber can rate the movie on a five-point ordinal scale. Based on its accumulated records of such ratings, Netflix gives the subscriber a predicted rating for any movie that that subscriber could choose to watch.

In 2006 Netflix announced a competition for developing an algorithm for predicting movie reviews. The winner would be awarded a \$1 million prize if the solution provided at least a 10% improvement in predictions, in terms of a particular root mean squared metric, over the algorithm then in use by Netflix. The training set consisted of about 100 million evaluations on 18,000 movies made by almost 500,000 Netflix subscribers. The average number of ratings per subscriber was 208, a very small fraction of the possible movies for evaluation.

For simplicity here, we'll imagine a binary positive versus negative rating rather than a five-point rating. One way to portray the data set is then as a $18,000 \times 500,000$ movie-by-subscriber matrix, in which a binary rating is shown in 100 million of the cells and the other cells have no data. Another portrayal of the data, but impractical, is as a $3^{18,000}$ cell contingency table that cross-classifies ratings using categories (positive, negative, unrated) on the 18,000 movies.

For a given subscriber for whom you have evaluations on some subset of the movies, how do you predict that subscriber's rating on some other movie? It's unclear how to do this with standard modeling methods. Logistic regression is not readily relevant, as no other subscriber may have seen the same movies as well as the one to be rated. Using existing ratings, we could measure the similarity between subscribers and/or the similarity between movies and then apply a nearest neighbor smoothing approach. With subscribers as the units, we predict the movie rating by averaging ratings for that movie by subscribers with similar opinions on jointly-watched movies. With movies as the units, we predict the movie rating by averaging ratings by that subscriber for similar movies. Although simple, such an approach depends on a choice of similarity metric, there may be few or no close neighbors for some movies, and many highly correlated neighbors for a movie may result in that set receiving too much weight.

An alternative approach, described by the team that won the Netflix competition (Bell et al. 2010), used latent variables. Both subscribers and movies were summarized on a latent vector of much smaller dimension, where the components of the vector referred to characteristics such as amount of violence, the amount of drama versus comedy, independent versus large-budget Hollywood star-driven movies, and characteristics that might not have a ready interpretation. A subscriber's rating was based on the inner product of the values of the subscriber's latent vector and the movie's latent vector. The model was fitted with L_2 -norm penalized likelihood methods. Extensions of the method used the fact that the set of movies a subscriber chooses to rate is an additional source of information about that person's tastes, and each subscriber's parameters could gradually vary over time.

Bell et al. (2010) explained that an ensemble method that generates multiple predictions from a variety of methods and then averages over them tends to generate better predictions than any one method. For the Netflix prize, this ensemble method incorporated both latent variable and nearest neighbor models.

7.5.8 Example: Credit Scoring

Credit scoring is the term describing methods for classifying applicants for credit into “good” and “bad” risk classes. According to Hand and Henley (1997), many credit scoring databases have more than 100,000 applicants measured on more than 100 variables. The probability that an applicant will default must be estimated based on characteristics such as annual income, occupation, marital status, age, post code, credit card possession, length of time at current address, type of bank account, court judgments, time with employer, time with bank, and details of loan payments.

Here, although p is large, n is much larger, so the challenges are not as severe as in cases where p exceeds n . Methods used include logistic regression, nearest neighbor smoothing, and non-model-based classification methods presented in Chapter 15, such as linear discriminant analysis. A complication is lots of missing values, although this itself can be a useful indicator for classification. Inherently continuous variables such as income are measured with discrete categories, and expert knowledge may impose monotonicity constraints on effects for different levels of a factor (e.g., such that the probability of default decreases as a function of income, adjusting for other variables). When the risk classes are not well separated, the response probability is a rather flat function and there is the danger of overfitting, so penalized likelihood and other smoothing mechanisms can be useful. The sample is usually far from random, which limits possible statistical inference.

A goal in selecting predictors is to include ones that discriminate well between default and nondefault outcomes. The success in identifying default cases can be summarized with standard tools such as classification tables and ROC curves, plotting the true positive rate against the false positive rate for various probability thresholds for predicting default.

Other business-related applications can have, by contrast, enormous values for p as well as n . Examples are market basket data and website browsing behavior, as described at the beginning of Section 15.3.

NOTES

Section 7.1: Probit and Complementary Log–Log Models

7.1 Probits/log–log: Finney (1971) is a standard reference on probit modeling. Ashford and Sowden (1970) generalized the probit model for multivariate binary responses; see also Lesaffre and Molenberghs (1991) and Ochi and Prentice (1984). Wedderburn (1976) showed that the log likelihood function is concave for probit and complementary log–log links.

7.2 Utility/extreme-value/logit/probit: For the utility model in Section 7.1.1, suppose ε_y are independent extreme-value random variables (instead of normal), with cdf $F(\varepsilon) = \exp[-\exp(-\varepsilon)]$. Then, McFadden (1974) showed that $P(Y = 1)$ satisfies the logistic regression model, because the difference between two extreme-value random variables has the logistic distribution. See also Amemiya (1981) and Maddala (1983, p. 60). Chambers and Cox (1967) showed that it is difficult to distinguish between models using probit and logit links unless n is extremely large.

7.3 Other link functions: Other link functions have been proposed for binary data, including the inverse of the cdf of a t distribution (for which the probit is the limiting case as $df \rightarrow \infty$), a log-gamma link (Genter and Farewell 1985) for which probit, complementary log–log, and log–log are special cases, and a weighted average of logit, log–log, and complementary log–log links (Lang 1999). Prentice (1976b) and Stukel (1988) extended the scope of logistic regression by introducing shape parameters that modify the behavior of the curve in extreme probability regions and allow for asymmetric treatment of the two tails. Prentice (1975, 1976b) used the inverse cdf of the logarithm of an F random variable, for which $df_1 = df_2 = 2$ gives the logistic. Guerrero and Johnson (1982) applied the Box–Cox power transformation to the odds, for which the logit is a special case. For other generalizations, see Aranda-Ordaz (1981), Kateri and Agresti (2010), and Pregibon (1980).

Section 7.2: Bayesian Inference for Binary Regression

7.4 Bayes literature: Racine et al. (1986) used Bayesian methods to obtain a posterior interval for LD₅₀ for the probit model. Chaloner and Larntz (1989) used Bayesian methods to determine optimal experimental design for logistic regression. For other Bayesian work on case–control studies, see Li and Conti (2009), Mukherjee and Chatterjee (2008), and Müller and Roeder (1997). For Bayesian item response modeling, Tsutakawa and Lin (1986) specified prior distributions on response probabilities and used them to induce priors on model parameters, the approach extended to binary regression models by Bedrick et al. (1997) and Christensen et al. (2010). Ghosh and Mukherjee (2010) surveyed Bayesian work on item response modeling. For binary regression, Zellner and Rossi (1984) used Monte Carlo methods with importance sampling, giving particular attention to multivariate normal priors, Chen et al. (2008) and Ibrahim and Laud (1991) used the Jeffreys prior, and Wong and Mason (1985) considered multilevel models. Dey et al. (2000) edited a book on Bayesian analyses for GLMs. The 2010 text *Frontiers of Statistical Decision Making and Bayesian Analysis* in honor of James Berger contains a chapter on “Bayesian Categorical Data Analysis” that has separate contributions on smoothing (by J. Albert), on matched-pairs binary data (by M. Ghosh and B. Mukherjee), and on the choice of link functions for binary data (by M.-H. Chen and colleagues). Agresti and Hitchcock (2005), Congdon (2005), Leonard (1999), and Leonard and Hsu (1994) surveyed Bayesian methods for categorical data.

Section 7.3: Conditional Logistic Regression

7.3 Conditional logistic and exact: For more details about conditional logistic regression, see Section 11.2, Breslow (1976), Breslow and Powers (1978), Breslow et al. (1978), Breslow and

Day (1980, Chap. 7), Cox (1970), Farewell (1979), Hosmer and Lemeshow (2000, Chap. 5), Lloyd (1999, Chap. 7), Prentice (1976a), and Prentice and Breslow (1978). Liang (1984) showed that conditional ML estimators and conditional score tests are asymptotically equivalent to their unconditional counterparts under sampling from exponential families. For more on exact inference using conditional distributions for contingency tables and logistic regression, see Sections 16.5 and 16.6, Agresti (1992), Hirji et al. (1987), Mehta and Patel (1995), and the StatXact and LogXact manuals (Cytel Software). Mehta et al. (2000) discussed Monte Carlo approximations. For improved higher-order asymptotic methods, see Brazzale and Davison (2008), Brazzale et al. (2007), and Davison et al. (2006).

Section 7.4: Smoothing: Kernels, Penalized Likelihood, Generalized Additive Models

7.6 Nearest neighbors and other kernels: See Hastie et al. (2009, Chap. 13) and references therein for details about the nearest neighbor method. Natural application areas are ones for which the data occur in physical space. See Besag (1974). Smoothing methods for binary data extend to multinomial responses. See Aitchison and Aitken (1976) and Exercise 7.32 for a simple kernel smoother. Hall and Titterington (1987) studied rates of convergence for multinomial kernel estimators and defined one that achieves the optimal rate. Ordinary kernel estimators tend to be biased toward zero at the boundary of a table. Dong and Simonoff (1994) dealt with improving kernel estimates on the boundary of large, sparse contingency tables.

7.6 Penalized likelihood, GAMs, smoothing surveys: Good and Gaskins (1971) introduced penalized likelihood methods. For more about the lasso, see Hastie et al. (2009, Sec. 3.4, 3.8). Simonoff (1983, 1996, 1998) proposed penalized likelihood methods for multinomial data. Kauermann and Tutz (2001) proposed likelihood-ratio goodness-of-fit tests of GLMs and GAMs against smooth alternatives. Yee and Wild (1996) defined generalized additive models for nominal and ordinal responses. See also Hastie and Tibshirani (1990) and Tutz (2011, Sec. 10.3.2). For surveys of smoothing methods, see Fahrmeir and Tutz (2001, Chap. 5), Lloyd (1999, Chap. 5), Simonoff (1996, Chap. 6; 1998), and Tutz (2011, Chap. 6, 10). See Albert (2010) for Bayesian smoothing methods.

Section 7.5: Issues in Analyzing High-Dimensional Categorical Data

7.8 Regularization with large p : Fan and Lv (2010) and Tutz (2011) reviewed penalized likelihood methods for variable selection in high dimensions. Genkin et al. (2007) proposed a type of Bayesian lasso for logistic regression. Meier et al. (2008) extended the lasso to do variable selection on predefined groups of variables and suggested a penalty term that is intermediate between a lasso and a quadratic penalty.

7.9 Multiple testing: For variable selection procedures, Westfall and Wolfinger (1997) and Westfall and Young (1993) presented ways to adjust P -values to take multiple tests into account, the first reference focusing on discrete distributions. Dudoit et al. (2003) and Farcomeni (2008) surveyed the issues in large-scale multiple hypothesis testing. Benjamini and Hochberg (1995) noted that their approach corresponds to a constrained maximization problem that chooses a level α^* for each test that maximizes the number of rejections n_{+2} subject to the constraint $\alpha^* g/n_{+2} \leq \alpha$.

EXERCISES

Applications

7.1 Refer to Exercise 5.2 on cancer remission with labeling index (LI) as predictor. [Table 7.7](#) shows output for fitting a probit model. Interpret the parameter estimates (**a**) using characteristics of the normal cdf response curve, (**b**) finding the estimated rate of change in the probability of remission where it equals 0.50, (**c**) finding the difference between the estimated probabilities of remission at the upper and lower quartiles of LI, 28 and 14, and (**d**) describing the effect of LI on an underlying latent variable for remission.

[Table 7.7](#) Output for Exercise 7.1 on Probit Model for Cancer Remission

Parameter	Estimate	Error	Likelihood			
			Standard	Ratio	95%	
					Confidence	Limits
Intercept	-2.3178	0.7795	-4.0114	-0.9084	8.84	0.0029
LI	0.0878	0.0328	0.0275	0.1575	7.19	0.0073

7.2 For the horseshoe crab data ([Table 4.3](#)), fit a probit model to describe the effects of width and color as a factor on the probability of a satellite. Interpret effects and conduct inference.

7.3 For the flour beetle mortality data in [Table 7.1](#), fit models using the (**a**) logit, (**b**) probit, (**c**) complementary log–log, and (**d**) log–log links, with dosage entered in the model in standardized form. Report and interpret model parameter estimates. What aspect of the data pattern causes the model with log–log link to fit so poorly?

7.4 For [Table 5.3](#) on maternal alcohol consumption and child’s congenital malformations, report the posterior mean and standard deviation and a 95% posterior interval (equal-tail or HPD) for β in the linear logit model with scores (0, 0.5, 1.5, 4.0, 7.0). Use (**a**) the $N(0, 1000^2)$ prior on model parameters and (**b**) the $N(0, 1)$ prior. Compare results to those obtained with ML.

7.5 Refer to the previous exercise. Conduct Bayesian analyses with the probit link, using prior distributions that (on the probit scale) are comparable to the priors used for the logit link. Compare results to those obtained with ML probit and with the Bayesian logistic analysis.

7.6 For the horseshoe crab data ([Table 4.3](#)) available at the text website, conduct a Bayesian analysis for the logistic model with width and dark color indicator as predictors of the probability of satellites, using relatively uninformative normal priors. Interpret results, and compare them to the ML fit.

7.7 Refer to the example on endometrial cancer in Section 7.2.2. To obtain analyses for the corresponding probit model that give similar substantive results, how would you need to change σ from the values of 10 and 1 used in the logistic analyses? Conduct such analyses, and report and interpret the posterior mean and standard deviation and a 95% equal-tail or HPD posterior interval for β_1 .

7.8 In Exercise 6.20, the main effects logistic model had all ML estimates infinite. By contrast, conduct a Bayesian analysis with independent $N(0, \sigma^2)$ priors. Show how the posterior mean and standard deviation estimates of the effects compare for $\sigma = 1$, $\sigma = 10$, and $\sigma = 100$.

7.9 Construct an artificial example of binary data with a single quantitative predictor and select a normal prior such that Bayesian inference gives substantively different results than frequentist inference. Discuss the factors that cause the results to be so different.

7.10 Construct a 2×2 table for a binary response and binary predictor x such that analyses you conduct for a logistic or probit model with linear predictor $\alpha + \beta x$ have ML estimate $\hat{\beta} = \infty$ but Bayesian posterior mean and posterior intervals for β that are finite.

7.11 For the 2×2 table with counts (by row) of (3, 1 / 1, 3), conduct a Bayesian analysis for the model, $\text{logit}[P(Y=1)] = \alpha + \beta x$, using $N(0, 1)$ priors for α and β .

- a.** With x coded as (1, 0), find the posterior mean and standard deviation for β .

b. With x coded as $(0.5, -0.5)$, find the posterior mean and standard deviation for β . Compare to (a), and explain why results differ somewhat. Would different results happen for β and its SE with frequentist analyses for these different codings of x ?

7.12 For the $2 \times 2 \times 5$ [Table 6.16](#), the small cell counts make large-sample analyses questionable. Conduct a small-sample test of conditional independence, and interpret.

7.13 [Table 7.8](#) comes from a 1987 study of nonmetastatic osteosarcoma (A. Goorin, *J. Clin. Oncol.* **5**: 1178–1184, and *LogXact* manual). The response is whether the subject achieved a three-year disease-free interval.

Table 7.8 Data for Exercise 7.13

Lymphocytic Infiltration	Gender	Osteoblastic Pathology	Disease-Free	
			Yes	No
High	Female	No	3	0
		Yes	2	0
	Male	No	4	0
		Yes	1	0
Low	Female	No	5	0
		Yes	3	2
	Male	No	5	4
		Yes	6	11

Source: *LogXact 7*. Cambridge, MA: CYTEL Software, 2005, p. 171.

- a.** Show that each predictor has a significant effect when used individually without the others.
- b.** Try to fit a main-effects logistic regression model containing all three predictors. Explain why the ML estimate for the effect of lymphocytic infiltration is infinite.
- c.** Using conditional logistic regression, conduct an exact test for the effect of lymphocytic infiltration, controlling for the other variables. Interpret results.

7.14 Using formula [\(7.13\)](#), smooth the data in Exercise 5.22 relating income to having a travel credit card. Graph results for three values of the smoothing parameter, corresponding to ones that (in your opinion) smooth too much, too little, and about the right amount.

7.15 Use the Firth penalized likelihood method to obtain finite estimates for the data in [Figure 6.5](#). Compare the 95% profile penalized likelihood confidence interval for β to the corresponding ordinary profile likelihood interval.

7.16 Using a generalized additive model, construct a figure like [Figure 5.2](#) for the horseshoe crab data, but using weight instead of width as the predictor.

7.17 Smooth the count data in [Figure 4.4](#) using generalized additive models. Graph results for three values of the smoothing parameter, corresponding to ones that (in your opinion) smooth too much, too little, and about the right amount.

7.18 The credit-scoring data file at www.statistik.lmu.de/service/datenarchiv/kredit/kredit_e.html includes 20 covariates for 1000 observations. Build a model for credit-worthiness, using as potential predictors: running account, duration of credit, payment of previous credits, intended use, gender, marital status.

7.19 Project: Go to a site with large data files, such as the UCI Machine Learning Repository (archive.ics.uci.edu/ml). Find a data set of interest to you that has a binary response variable. Use at least one method discussed in this chapter to learn something about the data. Summarize your analyses in a two-page report, attaching an appendix showing your use of software.

Theory and Methods

7.20 Refer to the *threshold model* used in Section 7.1.1 to motivate the probit model.

- a.** For identifiability, explain why you can set $\sigma = 1$ and $\tau = 0$. Explain why β then represents the expected number of standard deviation change in Y^* for a 1-unit increase in x .

b. Suppose you fitted this model separately to each of two groups and wanted to compare the effects of x for those groups. Suppose that the two groups had different residual variability for their underlying latent variable. Explain why even if the β parameters are identical for the two groups for the latent model, the corresponding effect parameters are not the same for the probit (or logistic) models actually used. [Allison (1999) discussed this issue and proposed an alternative way of comparing coefficients that can adjust for unequal residual variation.]

7.21 For independent binary $\{y_i\}$, from scratch (without using any results for GLMs) show that the likelihood equations for the logistic and probit regression models are

$$\sum_i (y_i - \hat{\pi}_i) z_i x_{ij} = 0, \quad j = 0, \dots, p,$$

where $z_i = 1$ for the logistic case and $z_i = \phi(\sum_j \hat{\beta}_j x_{ij})/\hat{\pi}_i(1 - \hat{\pi}_i)$ for the probit.

7.22 For the linear probability model $\pi_i = \alpha + \beta x_i$ applied with independent binary $\{y_i\}$, show that the likelihood equations are

$$\sum_i \left(\frac{y_i}{\pi_i} - \frac{1 - y_i}{1 - \pi_i} \right) = 0, \quad \sum_i \left(\frac{y_i}{\pi_i} - \frac{1 - y_i}{1 - \pi_i} \right) x_i = 0.$$

7.23 Derive the estimated asymptotic covariance matrix of $\hat{\beta}$ for the probit model from the GLM expression (4.31). [Hint: Recall that in the binomial case in Section 4.4, y_i is the *proportion* of successes.]

7.24 Consider model (7.3) with complementary log-log link.

a. Find x at which $\pi(x) = \frac{1}{2}$.

b. Show the greatest rate of change of $\pi(x)$ occurs at $x = -\alpha/\beta$. What does $\pi(x)$ equal at that point? Give the corresponding result for the model with log-log link, and compare to the logistic and probit models.

7.25 For the log-log model (7.4), explain how to interpret β .

7.26 Find the likelihood equations and apply (4.31) to find the form of the asymptotic covariance matrix of $\hat{\beta}$ for a binary GLM using the complementary log-log link function.

7.27 In logistic regression, suppose you use a Bayesian approach with an uninformative prior such as $N(0, 1000^2)$ for each model parameter. For any particular setting of the explanatory variables, explain why this implies that nearly all the prior weight is placed on probability values very close to 0 and very close to 1.

7.28 In a binary regression model, one of the explanatory variables is binary. For Bayesian fitting of the model, what is the reason for using coding such as $(0.5, -0.5)$ or $(1, -1)$ for levels of the binary predictor, instead of the usual $(1, 0)$ indicator coding?

7.29 For interval estimation of a logistic regression model parameter β_j , explain why the Bayesian highest posterior density interval is appropriate for β_j but not for the odds ratio effect $\exp(\beta_j)$. [Hint: See Section 3.6.5.]

7.30 For independent binomial sampling with the model $\text{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z$ for a $2 \times 2 \times K$ table, construct the log likelihood and identify the sufficient statistics to be conditioned out to perform exact inference about β .

7.31 Refer to the kernel smoother (7.12). Show that $\sum_i \bar{\pi}_i = 1$ if and only if the column totals of K equal 1.

7.32 For a multinomial distribution with c unordered categories, Aitchison and Aitken (1976) proposed a kernel estimator of form (7.12), having

$$k_{ij} = \gamma, \quad i = j \\ = (1 - \gamma)/(c - 1), \quad i \neq j$$

for $(1/c) \leq \gamma \leq 1$.

a. Show that the resulting kernel estimator of π has form

$$(1 - \lambda)\mathbf{p} + \lambda(\mathbf{1}/c),$$

where $\lambda = c(1 - \gamma)/(c - 1)$, which shrinks the sample proportions toward $(1/c, \dots, 1/c)$.

b. Show that as γ decreases from 1 to $1/c$, λ increases from 0 to 1. [Brown and Rundell (1985) proved that when no $\pi_i = 1$, a λ value exists such that the total mean squared error is smaller for this kernel estimator than for the sample proportions.]

c. Show that the kernel estimator of form in (a) is the same as the Bayes estimator (1.19) for the Dirichlet prior with $\{\alpha_i = \lambda n/(1 - \lambda)c\}$. Using this result, suggest a way of letting the data determine the value of λ in the kernel estimator.

7.33 Refer to Copas's kernel smoother (7.13) for binary regression, with $\phi(u) = \exp(-u^2/2)$.

a. To describe how close this estimator falls at a particular x value to a corresponding smoothing in the population, use the delta method to show that an estimated asymptotic variance is

$$\tilde{\pi}(x)[1 - \tilde{\pi}(x)] \frac{\sum_i \phi[\sqrt{2}(x - x_i)/\lambda]}{\{\sum_i \phi[(x - x_i)/\lambda]\}^2}.$$

Explain why this decreases as λ increases, and explain the implication.

b. As λ increases unboundedly, explain intuitively why $\tilde{\pi}(x)$ converges to $p = (\sum_i y_i)/n$ and this estimated asymptotic variance is approximately $p(1 - p)/n$.

7.34 Use a probabilistic argument to prove that the Bonferroni method works.

7.35 In Table 7.6, explain why (a) $P(n_{12} > 0)$ is a *familywise error rate* (FWER), (b) $E(n_{12})$ is a *per-family error rate* (PEER), and (c) $E(n_{12})/g$ is a *per-comparison error rate* (PCER).

7.36 Refer to the previous exercise.

a. Explain why multiple testing procedures satisfy $PCER \leq FWER \leq PFER$. Explain why for a fixed level for type I error rates, a procedure that controls the PFER is most conservative, leading to the fewest rejections of null hypotheses.

b. If hypothesis H_j is tested at level $\alpha_j, j = 1, \dots, g$, then under the complete null hypothesis, explain why (Dudoit et al. 2003)

$$PCER = \frac{\sum_j \alpha_j}{g} \leq \max(\alpha_1, \dots, \alpha_g) \leq FWER \leq PFER = \sum_j \alpha_j.$$

7.37 Read one of the genomics papers cited in Section 7.5.5 and prepare a two-page report summarizing its main contributions.

¹For example, for a single binomial parameter π , an improper uniform density for $\text{logit}(\pi)$ corresponds to an improper beta prior for π with $\alpha_1 = \alpha_2 = 0$.

²As of 2012, see www.ics.uci.edu/~wjohnson/BIDA/Ch8/trauma300.txt.

³This is (16.29) in Chapter 16, setting $\theta = 1$.

⁴For example *plink* for whole genome association analysis, described at pngu.mgh.harvard.edu/~purcell/plink.

⁵Figure 1 in Fan and Lv (2010) illustrates these issues.

⁶See P. Kraft and D. J. Hunter, "Genetic risk prediction—Are we there yet," *N. Engl. J. Med.* **360**: 1701–1703, 2009; and K. Wang et al., "Analyzing biological pathways in genome-wide association studies," *Nature Rev. Genetics* **11**: 843–854, 2010.

CHAPTER 8

Models for Multinomial Responses

In Chapters 5, 6, and 7, we modeled binary response variables with *binomial* GLMs. Multicategory responses use *multinomial* GLMs. In this chapter we generalize logistic regression to handle multinomial response variables, with separate models for nominal and ordinal cases.

In Section 8.1 we present a model for nominal responses. It uses a separate binary logistic equation for each pair of response categories. In Section 8.2 we present a model for ordinal responses, using logits of cumulative response probabilities. In Section 8.3 we use other link functions for those cumulative probabilities and consider alternative ordinal logit models.

In Section 8.4 we present tests of conditional independence with multinomial responses using models and using generalizations of the Cochran–Mantel–Haenszel statistic. In Section 8.5 we introduce a multinomial logit model for *discrete-choice modeling* of a subject’s choice from one of several options when values of predictors may depend on the option. The final section discusses Bayesian methods for multinomial response modeling.

8.1 NOMINAL RESPONSES: BASELINE-CATEGORY LOGIT MODELS

For a nominal-scale response variable Y with J categories, multcategory (also called *polytomous*) logistic models simultaneously describe the log odds for all $\binom{J}{2}$ pairs of categories. Given a certain choice of $J - 1$ of these, the rest are redundant.

8.1.1 Baseline-Category Logits

Let $\pi_j(\mathbf{x}) = P(Y = j|\mathbf{x})$ at a fixed setting \mathbf{x} for explanatory variables, with $\sum_j \pi_j(\mathbf{x}) = 1$. For observations at that setting, we treat the counts at the J categories of Y as a multinomial variate with probabilities $\{\pi_1(\mathbf{x}), \dots, \pi_J(\mathbf{x})\}$. Logistic models pair each response category with a baseline category, such as the last one or the most common one. Consider the model

$$(8.1) \quad \log \frac{\pi_j(\mathbf{x})}{\pi_J(\mathbf{x})} = \alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}, \quad j = 1, \dots, J - 1.$$

The left-hand side is the logit of a conditional probability, $\text{logit}[P(Y=j|Y=j \text{ or } Y=J)]$. This model simultaneously describes the effects of \mathbf{x} on these $J - 1$ logits. The effects vary according to the response paired with the baseline. These $J - 1$ equations determine parameters for logits with other pairs of response categories, since

$$\log \frac{\pi_a(\mathbf{x})}{\pi_b(\mathbf{x})} = \log \frac{\pi_a(\mathbf{x})}{\pi_J(\mathbf{x})} - \log \frac{\pi_b(\mathbf{x})}{\pi_J(\mathbf{x})}.$$

With categorical predictors, X^2 and G^2 goodness-of-fit statistics provide a model check when data are not sparse. When an explanatory variable is continuous or the data are sparse, such statistics are valid only for comparing nested models differing by relatively few terms.

8.1.2 Example: Alligator Food Choice

[Table 8.1](#) is from a study of factors influencing the primary food choice of alligators. The study captured 219 alligators in four Florida lakes. The nominal response variable is the primary food type, in volume, found in an alligator's stomach. This had five categories: fish, invertebrate, reptile, bird, other. The invertebrates included apple snails, aquatic insects, and crayfish. The reptiles were primarily turtles, although one stomach contained the tags of 23 baby alligators released in the lake the previous year! The “other” category consisted of amphibian, mammal, plant material, stones or other debris, or no food or dominant type. [Table 8.1](#) also classifies the alligators according to L = lake of capture (Hancock, Oklawaha, Trafford, George), G = gender (male, female), and S = size (≤ 2.3 meters long, > 2.3 meters long).

Table 8.1 Primary Food Choice of Alligators, by Lake, Gender, and Size of the Alligator

Lake	Gender	Size (m)	Primary Food Choice				
			Fish	Invertebrate	Reptile	Bird	Other
Hancock	Male	≤ 2.3	7	1	0	0	5
		> 2.3	4	0	0	1	2
	Female	≤ 2.3	16	3	2	2	3
		> 2.3	3	0	1	2	3
Oklawaha	Male	≤ 2.3	2	2	0	0	1
		> 2.3	13	7	6	0	0
	Female	≤ 2.3	3	9	1	0	2
		> 2.3	0	1	0	1	0
Trafford	Male	≤ 2.3	3	7	1	0	1
		> 2.3	8	6	6	3	5
	Female	≤ 2.3	2	4	1	1	4
		> 2.3	0	1	0	0	0
George	Male	≤ 2.3	13	10	0	2	2
		> 2.3	9	0	0	1	2
	Female	≤ 2.3	3	9	1	0	1
		> 2.3	8	1	0	0	1

Source: Data courtesy of Clint Moore, from an unpublished manuscript by M. F. Delaney and C. T. Moore.

Baseline-category logit models can investigate the effects of L , G , and S on primary food type. [Table 8.2](#) contains fit statistics for several models. We denote a model by its predictors: for instance, $(L + S)$ has additive lake and size effects, and $()$ has no predictors. The data are sparse, 219 observations scattered among 80 cells. Thus, G^2 is more reliable for comparing models than for testing a model's fit. The statistics $G^2[()|(G)] = 2.1$ and $G^2 = [(L + S)|(G + L + S)] = 2.2$, each based on $df = 4$, suggest simplifying by collapsing the table over gender. (Other analyses, not presented here, show that adding interaction terms including G do not improve the fit significantly.) The G^2 and X^2 values for reduced models for the collapsed table indicate that both L and S have effects. [Table 8.3](#) exhibits fitted values for model $(L + S)$ for the collapsed table. Absolute values of standardized residuals comparing observed and fitted values exceed 2 in only two of the 40 cells and exceed 3 in none of the cells. The fit seems adequate.

Table 8.2 Goodness of Fit of Baseline-Category Logit Models for [Table 8.1](#) on Alligator Primary Food Choice

Model ^a	G^2	X^2	df	Collapsed over G	G^2	X^2	df
($)$	116.8	106.5	60	($)$	81.4	73.1	28
(G)	114.7	101.2	56				
(S)	101.6	86.9	56	(S)	66.2	54.3	24
(L)	73.6	79.6	48	(L)	38.2	32.7	16
($L + S$)	52.5	58.0	44	($L + S$)	17.1	15.0	12
($G + L + S$)	50.3	52.6	40				

^a G , gender; S , size; L , lake of capture.

Table 8.3 Observed and Fitted Values for Baseline-Category Logit Model Using Lake and Size

of Alligator Main Effects to Predict Primary Food Choice

Lake	Size of Alligator (meters)	Primary Food Choice				
		Fish	Invertebrate	Reptile	Bird	Other
Hancock	≤ 2.3	23 (20.9)	4 (3.6)	2 (1.9)	2 (2.7)	8 (9.9)
	> 2.3	7 (9.1)	0 (0.4)	1 (1.1)	3 (2.3)	5 (3.1)
Olkawaha	≤ 2.3	5 (5.2)	11 (12.0)	1 (1.5)	0 (0.2)	3 (1.1)
	> 2.3	13 (12.8)	8 (7.0)	6 (5.5)	1 (0.8)	0 (1.9)
Trafford	≤ 2.3	5 (4.4)	11 (12.4)	2 (2.1)	1 (0.9)	5 (4.2)
	> 2.3	8 (8.6)	7 (5.6)	6 (5.9)	3 (3.1)	5 (5.8)
George	≤ 2.3	16 (18.5)	19 (16.9)	1 (0.5)	2 (1.2)	3 (3.8)
	> 2.3	17 (14.5)	1 (3.1)	0 (0.5)	1 (1.8)	3 (2.2)

Fish was the most common food choice. We now estimate the effects of lake and size on the odds that alligators select other primary food types instead of fish. Let $s = 1$ for size ≤ 2.3 meters and 0 otherwise, let z_H be an indicator variable for Lake Hancock ($z_H = 1$ for alligators in that lake and 0 otherwise), and let z_O and z_T be indicator variables for Lakes Oklawaha and Trafford. With fish as the baseline category. [Table 8.4](#) contains ML estimates of effect parameters. We use letter subscripts to denote the food choice categories. For example, the prediction equation for the log odds of selecting invertebrates instead of fish is

Table 8.4 Estimated Parameters in Baseline-Category Logit Model for Alligator Food Choice, Based on Indicator Variable for Size (1 = Small, 0 = Large) and for Each Lake except Lake George^a

Logit ^b	Intercept	Size ≤ 2.3	Lake		
			Hancock	Olkawaha	Trafford
$\log(\pi_I/\pi_F)$	-1.55	1.46 (0.40)	-1.66 (0.61)	0.94 (0.47)	1.12 (0.49)
$\log(\pi_R/\pi_F)$	-3.31	-0.35 (0.58)	1.24 (1.19)	2.46 (1.12)	2.94 (1.12)
$\log(\pi_B/\pi_F)$	-2.09	-0.63 (0.64)	0.70 (0.78)	-0.65 (1.20)	1.09 (0.84)
$\log(\pi_O/\pi_F)$	-1.90	0.33 (0.45)	0.83 (0.56)	0.01 (0.78)	1.52 (0.62)

^a SE values in parentheses.

^b Response categories: I , invertebrate; R , reptile; B , bird; O , other; F , fish.

$$\log(\hat{\pi}_I/\hat{\pi}_F) = -1.55 + 1.46s - 1.66z_H + 0.94z_O + 1.12z_T.$$

Size of alligator has a noticeable effect. For a given lake, for small alligators the estimated odds that primary food choice was invertebrates instead of fish are $\exp(1.46) = 4.3$ times the estimated odds for large alligators; the Wald 95% confidence interval is $\exp[1.46 \pm 1.96(0.396)] = (2.0, 9.3)$. The lake effects indicate that the estimated odds that the primary food choice was invertebrates instead of fish are relatively higher at Lakes Trafford and Oklawaha and relatively lower at Lake Hancock than they are at Lake George.

The equations in [Table 8.4](#) determine those for other food-choice pairs. For instance, for the pair (invertebrate, other),

$$\begin{aligned} \log(\hat{\pi}_I/\hat{\pi}_O) &= \log(\hat{\pi}_I/\hat{\pi}_F) - \log(\hat{\pi}_O/\hat{\pi}_F) \\ &= (-1.55 + 1.46s - 1.66z_H + 0.94z_O + 1.12z_T) \\ &\quad - (-1.90 + 0.33s + 0.83z_H + 0.01z_O + 1.52z_T) \\ &= 0.35 + 1.13s - 2.48z_H + 0.93z_O - 0.39z_T. \end{aligned}$$

Viewing all these, we see that size has its greatest impact in terms of whether invertebrates rather than fish are the primary food choice.

8.1.3 Estimating Response Probabilities

The equation that expresses multinomial logistic models directly in terms of response probabilities $\{\pi_j(\mathbf{x})\}$ is

$$(8.2) \quad \pi_j(\mathbf{x}) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x})}{1 + \sum_{h=1}^{J-1} \exp(\alpha_h + \boldsymbol{\beta}_h^T \mathbf{x})}$$

with $\alpha_J = 0$ and $\boldsymbol{\beta}_J = \mathbf{0}$. This follows from (8.1), noting that (8.1) also holds with $j = J$ by setting $\alpha_J = 0$ and $\boldsymbol{\beta}_J = \mathbf{0}$. (The parameters also equal zero for a baseline category for identifiability reasons; see Exercise 8.31.) The denominator of (8.2) is the same for each j . The numerators for various j sum to the denominator, so $\sum_j \pi_j(\mathbf{x}) = 1$. For $J = 2$, this formula simplifies to the binary logistic regression probability formula (5.1).

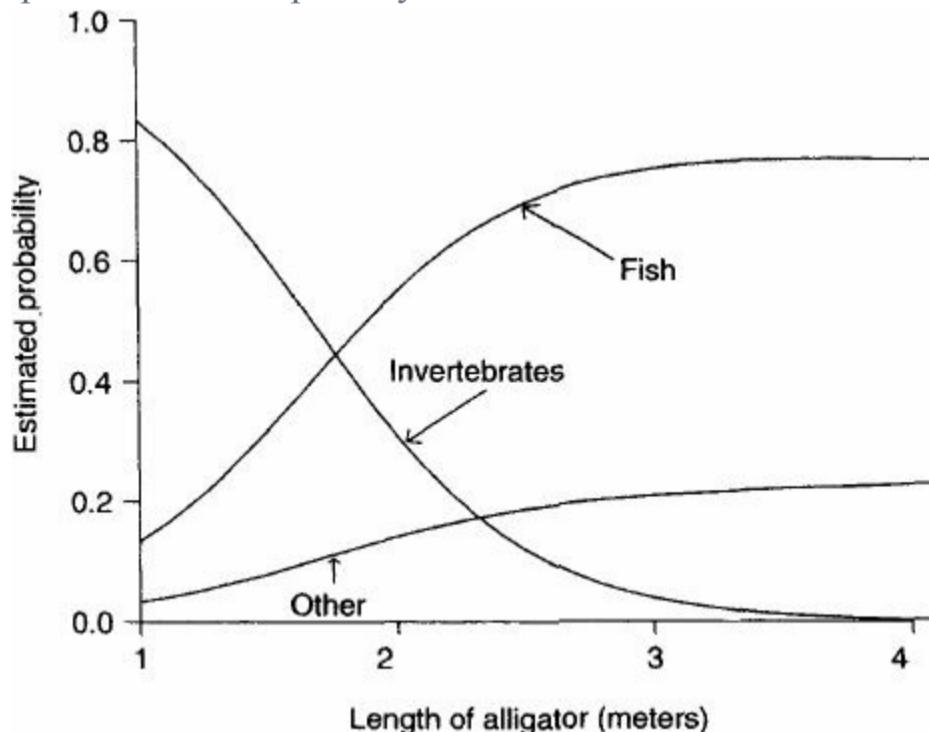
From Table 8.4 the estimated probability that a large alligator in Lake Hancock has invertebrates as the primary food choice is

$$\hat{\pi}_I = \frac{e^{-1.55-1.66}}{1 + e^{-1.55-1.66} + e^{-3.31+1.24} + e^{-2.09+0.70} + e^{-1.90+0.83}} = 0.023.$$

The estimated probabilities for (reptile, bird, other, fish) are (0.072, 0.141, 0.194, 0.570).

This example used qualitative predictors. Multinomial logit models can also contain quantitative predictors. In this study, the biologists used the size indicator variable to distinguish between adult and subadult alligators. However, the alligators' actual length was measured and is quantitative. With quantitative predictors, it is informative to plot the estimated probabilities. To illustrate, for alligators at one lake, Figure 8.1 plots the estimated probabilities that primary food choice is fish, invertebrate, or other (which combines the other, bird, and reptile categories) as a function of length. With more than two response categories, the probability for a given category need not continuously increase or decrease (Exercise 8.32).

Figure 8.1 Estimated probabilities for primary food choice.



8.1.4 Fitting Baseline-Category Logistic Models

ML fitting of multinomial logistic models maximizes the likelihood subject to $\{\pi_j(\mathbf{x})\}$ simultaneously satisfying the $J - 1$ equations that specify the model. Let $\mathbf{y}_i = (y_{i1}, \dots, y_{iJ})$ represent the multinomial trial for subject i , where $y_{ij} = 1$ when the response is in category j and $y_{ij} = 0$ otherwise, so $\sum_j y_{ij} = 1$. Let $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$ denote explanatory variable values for subject i . Let $\boldsymbol{\beta}_j = (\beta_{j1}, \dots, \beta_{jp})^T$ denote parameters for the j th baseline-category logit.

Since $\pi_J = 1 - (\pi_1 + \dots + \pi_{J-1})$ and $y_{iJ} = 1 - (y_{i1} + \dots + y_{i,J-1})$, the contribution to the log likelihood by subject i is

$$\begin{aligned}\log \left[\prod_{j=1}^J \pi_j(\mathbf{x}_i)^{y_{ij}} \right] &= \sum_{j=1}^{J-1} y_{ij} \log \pi_j(\mathbf{x}_i) + \left(1 - \sum_{j=1}^{J-1} y_{ij} \right) \log \left[1 - \sum_{j=1}^{J-1} \pi_j(\mathbf{x}_i) \right] \\ &= \sum_{j=1}^{J-1} y_{ij} \log \frac{\pi_j(\mathbf{x}_i)}{1 - \sum_{j=1}^{J-1} \pi_j(\mathbf{x}_i)} + \log \left[1 - \sum_{j=1}^{J-1} \pi_j(\mathbf{x}_i) \right].\end{aligned}$$

Thus, the baseline-category logits are the natural parameters for the multinomial distribution.

Next, we construct the likelihood equations, for n independent observations. In the last expression above, we substitute $\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}_i$ for the logit in the first term and

$$\pi_J(\mathbf{x}_i) = 1 / \left[1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}_i) \right]$$

in the second term. Then, the log likelihood is

$$\begin{aligned}\log \prod_{i=1}^n \left[\prod_{j=1}^J \pi_j(\mathbf{x}_i)^{y_{ij}} \right] &= \sum_{i=1}^n \left\{ \sum_{j=1}^{J-1} y_{ij}(\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}_i) - \log \left[1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}_i) \right] \right\} \\ &= \sum_{j=1}^{J-1} \left[\alpha_j \left(\sum_{i=1}^n y_{ij} \right) + \sum_{k=1}^p \beta_{jk} \left(\sum_{i=1}^n x_{ik} y_{ij} \right) \right] \\ &\quad - \sum_{i=1}^n \log \left[1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}_i) \right].\end{aligned}$$

The sufficient statistic for β_{jk} is $\sum_i x_{ik} y_{ij}$, $j = 1, \dots, J - 1$, $k = 1, \dots, p$. The sufficient statistic for α_j is $\sum_i y_{ij} = \sum_i x_{i0} y_{ij}$ for $x_{i0} = 1$; this is the total number of outcomes in category j .

The likelihood equations equate the sufficient statistics to their expected values. The log-likelihood function is concave, and the Newton–Raphson method yields the ML parameter estimates. The exception is when there is a choice of baseline category such that complete or quasi-complete separation occurs for each logit when paired with another category. In that case, some estimates and SE values are actually infinite.

The estimators have large-sample normal distributions. As usual, standard errors are square roots of diagonal elements of the inverse information matrix.

8.1.5 Multicategory Logit Model as a Multivariate GLM

For a univariate response variable in the natural exponential family, a GLM has form $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$ for a link function g , expected response $\mu_i = E(Y_i)$, vector of values \mathbf{x}_i of p explanatory variables for observation i , and parameter vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$. This extends to a *multivariate generalized linear model* for distributions in the multivariate exponential family (Exercise 8.29), such as the multinomial.

For response vector \mathbf{y}_i for subject i , with $\boldsymbol{\mu}_i = E(\mathbf{Y}_i)$, let g be a vector of link functions. The multivariate GLM has the form

$$(8.3) \quad \mathbf{g}(\boldsymbol{\mu}_i) = \mathbf{X}_i \boldsymbol{\beta},$$

where row h of the model matrix \mathbf{X}_i for observation i contains values of explanatory variables for y_{ih} (Fahrmeir and Tutz 2001, Chap. 3).

The baseline-category logit model is a multivariate GLM. Let $\mathbf{y}_i = (y_{i1}, \dots, y_{iJ-1})^T$, since y_{iJ} is redundant, $\boldsymbol{\mu}_i = (\pi_1(\mathbf{x}_i), \dots, \pi_{J-1}(\mathbf{x}_i))^T$ and

$$g_j(\mu_{ij}) = \log\{\mu_{ij}/[1 - (\mu_{i1} + \dots + \mu_{i,J-1})]\}.$$

The model matrix for observation i is

$$\mathbf{X}_i = \begin{pmatrix} 1 & \mathbf{x}_i^T \\ & 1 & \mathbf{x}_i^T \\ & & \dots \\ & & & 1 & \mathbf{x}_i^T \end{pmatrix}$$

with 0 entries in other locations, and $\boldsymbol{\beta}^T = (\alpha_1, \boldsymbol{\beta}_1^T, \dots, \alpha_{J-1}, \boldsymbol{\beta}_{J-1}^T)$.

8.1.6 Multinomial Probit Models

The multinomial logit model with baseline-category logits results from a latent utility representation that generalizes the one mentioned in Section 7.1.1. Let U_{ij} denote the utility of response outcome j for subject i . Suppose that

$$U_{ij} = \alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}_i + \epsilon_{ij}.$$

The response outcome for subject i is the value of j having maximum utility. McFadden (1974) showed that the assumption that $\{\epsilon_{ij}\}$ are independent and have the extreme value distribution (i.e., cdf $F(\epsilon) = \exp[-\exp(-\epsilon)]$) is equivalent to multinomial logit model (8.1) holding. The identifiable parameters for that model are $(\boldsymbol{\beta}_j - \boldsymbol{\beta}_J)$. Likewise, the utilities are identifiable in terms of relative utilities $(U_{ij} - U_{iJ})$.

It may seem more natural to assume that $\{\epsilon_{ij}\}$ have a normal distribution. Aitchison and Bennett (1970) suggested this approach, for independent standard normal variates. The corresponding model, called the *multinomial probit model*, gives a similar fit. For a particular explanatory variable x_k and pair of categories a and b , $(\beta_{ak} - \beta_{bk})$ describes the effect of a 1-unit increase in x_k on the difference between the mean utilities for those categories. If the normal distribution for $\{\epsilon_{ij}\}$ had instead been scaled to have some fixed standard deviation σ , then $(\beta_{ak} - \beta_{bk})$ would describe the difference in mean utilities in terms of the number of standard deviations of the utility distribution.

Fitting the multinomial probit model is computationally more complex than the corresponding logit model. Finding the likelihood function requires numerical integration, because

$$\begin{aligned} \pi_j(\mathbf{x}_i) &= P(U_{ij} > U_{ik}, \text{ for all } k \neq j) = E_{U_{ij}}[P(U_{ik} < u_{ij}, \text{ for all } k \neq j | U_{ij} = u_{ij})] \\ &= \int \phi(u_{ij} - \alpha_j - \boldsymbol{\beta}_j^T \mathbf{x}_i) \prod_{k \neq j} \Phi(u_{ij} - \alpha_k - \boldsymbol{\beta}_k^T \mathbf{x}_i) du_{ij}, \end{aligned}$$

for the standard normal pdf ϕ and cdf Φ .

It often seems unrealistic to expect the errors for different outcomes in the utility latent model to be uncorrelated. A more general model permits an arbitrary covariance matrix for $(\epsilon_{i1}, \dots, \epsilon_{iJ})$, with $\text{var}(\epsilon_{i1}) = 1$ for identifiability. Fitting is then even more complex. Natarajan et al. (2000) proposed a Monte Carlo EM algorithm for ML estimation that has the advantage of circumventing direct evaluation of the likelihood function by taking advantage of the latent structure. See also Imai and van Dyk (2005) and McCulloch et al. (2000), who utilized a corresponding latent variable model introduced in Section 8.6.3.

8.1.7 Example: Effect of Menu Pricing

Natarajan et al. (2000) described a study to investigate the effect of the pricing of a fish dish in a restaurant on a customer's choice among four popular food choices. On several winter Fridays or Saturdays the fish dish was priced between \$8.95 and \$10.95. Data were collected for 974 orders. Treating the fish dish as the baseline category, the multinomial probit model provided three equations for the difference between the predicted utility for each food item and fish.

For example, the equation relating steak (the first item) to the fish dish had predicted utility difference for subject i of

$$\hat{U}_{i1} - \hat{U}_{i4} = 0.168 - 0.502F_i - 0.072P_i,$$

where $F_i = 1$ for Friday and 0 for Saturday, and P_i is the price of the fish item when subject i ordered. The standard errors were 0.178 for the Friday effect and 0.072 for the fish pricing effect. So, the fish pricing did not have a significant effect on the choice between fish and steak (higher price even having a negative estimated effect on selecting steak). Natarajan et al. (2000) used a general covariance structure for the normal errors, with $\text{var}(U_{i1} - U_{i4}) = 1.0$ for identifiability. Thus, the estimated effect of Friday was to depress the utility for steak relative to fish by half a standard distribution of the normal distribution for the utility difference.

8.2 ORDINAL RESPONSES: CUMULATIVE LOGIT MODELS

We have discussed the benefits of utilizing the ordinality of a variable by focusing inferences on a single parameter (e.g., see Section 5.3.7). These benefits extend to models for ordinal responses. Models with terms that reflect ordinal characteristics such as monotone trend have improved model parsimony and power. In this section we introduce the most popular logistic model for ordinal responses.

8.2.1 Cumulative Logits

We utilize the category ordering by forming logits of cumulative probabilities,

$$P(Y \leq j|\mathbf{x}) = \pi_1(\mathbf{x}) + \cdots + \pi_j(\mathbf{x}), \quad j = 1, \dots, J.$$

The *cumulative logits* are defined as

$$\begin{aligned} \text{logit}[P(Y \leq j|\mathbf{x})] &= \log \frac{P(Y \leq j|\mathbf{x})}{1 - P(Y \leq j|\mathbf{x})} \\ (8.4) \quad &= \log \frac{\pi_1(\mathbf{x}) + \cdots + \pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x}) + \cdots + \pi_J(\mathbf{x})}, \quad j = 1, \dots, J-1. \end{aligned}$$

Each cumulative logit uses all J response categories.

8.2.2 Proportional Odds Form of Cumulative Logit Model

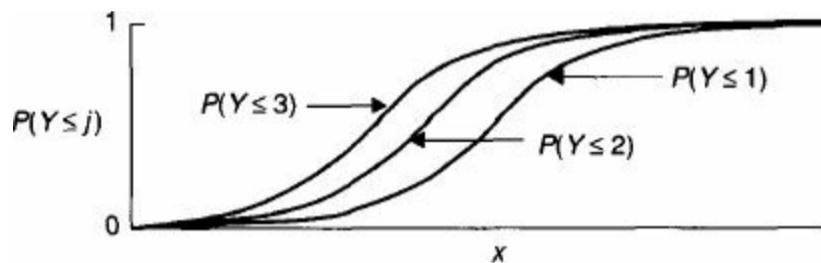
A model for $\text{logit}[P(Y \leq j)]$ alone is an ordinary logistic model for a binary response in which categories 1 to j form one outcome and categories $j + 1$ to J form the second. A model that simultaneously uses all $(J - 1)$ cumulative logits in a single parsimonious model is

$$(8.5) \quad \text{logit}[P(Y \leq j|x)] = \alpha_j + \beta^T x, \quad j = 1, \dots, J - 1.$$

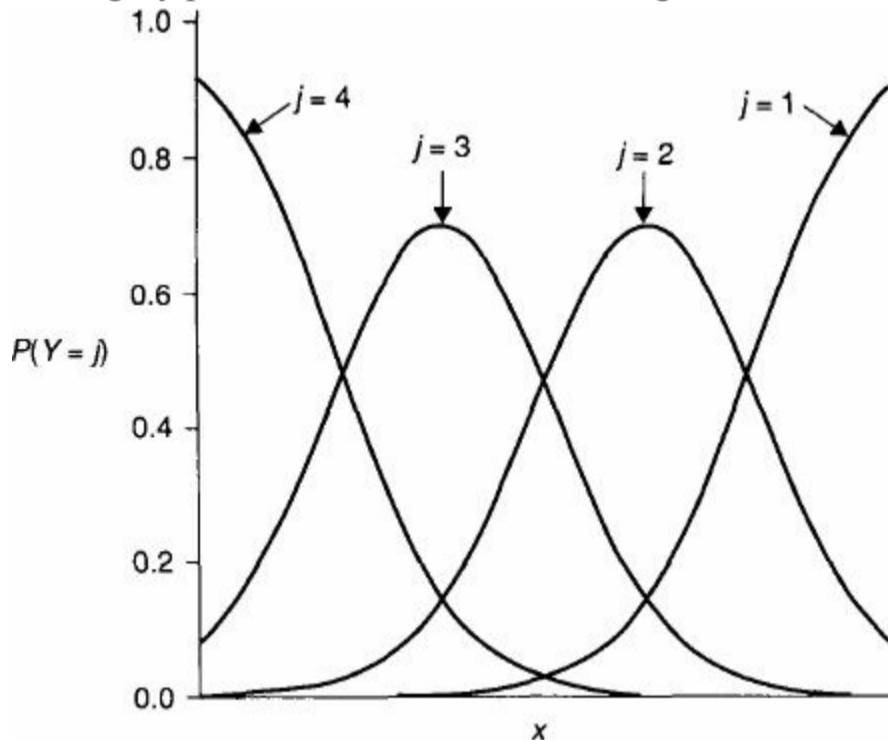
Each cumulative logit has its own intercept. The $\{\alpha_j\}$ are increasing in j , because $P(Y \leq j|x)$ increases in j for fixed x and the logit is an increasing function of $P(Y \leq j|x)$.

This model assumes the same effects β for each logit. For a single continuous predictor x , [Figure 8.2](#) depicts the model when $J = 4$. For fixed j , the response curve is a logistic regression curve for a binary response with outcomes $(Y \leq j)$ and $(Y > j)$. The curves for $j = 1, 2$, and 3 have the same shape. They share exactly the same rate of increase or decrease but are horizontally displaced from each other. [Figure 8.3](#) portrays the corresponding curves for the category probabilities.

[Figure 8.2](#) Cumulative logit model with the same effect on each of three cumulative probabilities in a four-category response.



[Figure 8.3](#) Individual category probabilities in cumulative logit model with four response categories.



The cumulative logit model [\(8.5\)](#) satisfies

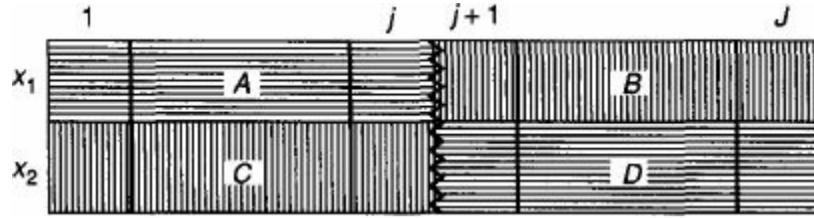
$$\begin{aligned} & \text{logit}[P(Y \leq j|x_1)] - \text{logit}[P(Y \leq j|x_2)] \\ &= \log \frac{P(Y \leq j|x_1)/P(Y > j|x_1)}{P(Y \leq j|x_2)/P(Y > j|x_2)} = \beta^T(x_1 - x_2). \end{aligned}$$

An odds ratio of cumulative probabilities is called a *cumulative odds ratio*. The odds of making response $\leq j$ at $x = x_1$ are $\exp[\beta^T(x_1 - x_2)]$ times the odds at $x = x_2$. The log cumulative odds ratio is proportional to the distance between x_1 and x_2 . The same proportionality constant applies to each logit. Because of this property, model [\(8.5\)](#) is often called a *proportional odds model* (McCullagh 1980).

With a single predictor, the cumulative odds ratio equals e^β whenever $x_1 - x_2 = 1$. [Figure 8.4](#) illustrates the constant cumulative odds ratio this model then implies for all j . It shows the J -category

response collapsed into the binary outcome ($\leq j$, $>j$) and shows the sets of cells that determine the cumulative odds ratio that takes the same value e^β for each such collapsing.

Figure 8.4 Uniform odds ratios AD/BC whenever $x_1 - x_2 = 1$, for all binary collapsings of the response in cumulative logit model of proportional odds form.



Model (8.5) constrains the $J - 1$ response curves to have the same shape. For multicategory indicator (y_{i1}, \dots, y_{iJ}) of the response for subject i , the product multinomial likelihood function is

$$\begin{aligned}
 \prod_{i=1}^n \left[\prod_{j=1}^J \pi_j(\mathbf{x}_i)^{y_{ij}} \right] &= \prod_{i=1}^n \left\{ \prod_{j=1}^J [P(Y \leq j | \mathbf{x}_i) - P(Y \leq j-1 | \mathbf{x}_i)]^{y_{ij}} \right\} \\
 &= \prod_{i=1}^n \left\{ \prod_{j=1}^J \left[\frac{\exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_j + \boldsymbol{\beta}^T \mathbf{x}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}^T \mathbf{x}_i)} \right]^{y_{ij}} \right\},
 \end{aligned} \tag{8.6}$$

viewed as a function of $(\{\alpha_j\}, \boldsymbol{\beta})$. This can be maximized to obtain the ML estimates using Fisher scoring (McCullagh 1980, Walker and Duncan 1967) or the Newton-Raphson method. The SE values differ somewhat, as the expected information and observed information matrices are not the same for this non-canonical-link model.

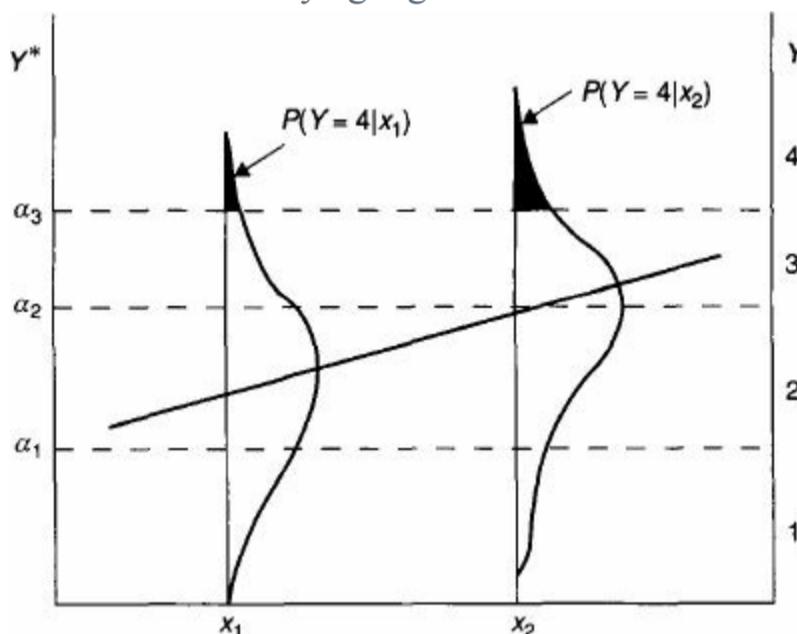
8.2.3 Latent Variable Motivation for Proportional Odds Structure

A regression model for a latent continuous variable assumed to underlie Y motivates the common effect β for different j in the proportional odds form of the model (Anderson and Philips 1981). Let Y^* denote this underlying latent variable. Suppose that it has cdf $G(y^* - \eta)$, where values of y^* vary around a location parameter η (such as a mean) that depends on x through $\eta(x) = \beta^T x$. Suppose that the thresholds $-\infty = \alpha_0 < \alpha_1 < \dots < \alpha_J = \infty$ are *cutpoints* of the continuous scale such that the observed response y satisfies

$$y = j \quad \text{if } \alpha_{j-1} < y^* \leq \alpha_j.$$

That is, y falls in category j when the latent variable falls in the j th interval of values, as [Figure 8.5](#) depicts. Then

[Figure 8.5](#) Ordinal measurement and underlying regression model for a latent variable.



$$P(Y \leq j|x) = P(Y^* \leq \alpha_j|x) = G(\alpha_j - \beta^T x).$$

The appropriate model for Y implies that the link function G^{-1} , the inverse of the cdf for Y^* , applies to $P(Y \leq j|x)$. If $Y^* = \beta^T x + \epsilon$, where the cdf G of ϵ is the standard logistic (Section 4.2.5), then G^{-1} is the logit link and a proportional odds model results. Normality for ϵ implies a probit link for cumulative probabilities (Section 8.3.2).

In this derivation, the same parameters β occur for the effects regardless of how the cutpoints $\{\alpha_j\}$ chop up the scale for the latent variable. The effect parameters are invariant to the choice of categories for Y . If a continuous variable measuring political ideology has a linear regression with some predictor variables, then the same β apply to a discrete version of political ideology with the categories (liberal, moderate, conservative) or (very liberal, slightly liberal, moderate, slightly conservative, very conservative). This feature makes it possible to compare estimates from studies using different response scales.

Using a cdf of form $G(y^* - \eta)$ for the latent variable resulted in linear predictor $\alpha_j - \beta^T x$ rather than $\alpha_j + \beta^T x$. When $\beta_k > 0$, as x_{ik} increases each cumulative logit then decreases, so each cumulative probability decreases and relatively less probability mass falls at the low end of the Y scale. Thus, Y tends to be larger at higher values of x_{ik} . With this parameterization the sign of β_k has the usual meaning. However, some software (e.g., SAS) uses form [\(8.5\)](#).

8.2.4 Example: Happiness and Traumatic Events

[Table 8.5](#) shows GSS data on $Y =$ happiness (categories 1 = very happy, 2 = pretty happy, 3 = not too happy), x_1 = total number of traumatic events that happened to the respondent and his/her relatives in the last year, and x_2 = race (1 = black, 0 = white). We restricted the age range to 18-22 in order to have a relatively small sample ($n = 97$), to illustrate how certain models may then have infinite ML estimates. In particular, only 13 of the 97 observations were in the black category of race, and of them, none had response in the very happy category. [Table 8.5](#) shows the data for four of the subjects. The complete data set is at the text website.

Table 8.5 Four Observations from Data Set on Happiness, Number of Traumatic Events, and Race

Observation	Happiness	Number of Traumatic Events	Race
1	Pretty happy	2	White
2	Pretty happy	3	Black
3	Very happy	0	White
4	Not too happy	5	White

Source: 1984 General Social Survey; complete data at www.stat.ufl.edu/~aa/cda/cda.html.

The main-effects cumulative logit model of proportional odds form [\(8.5\)](#) is

$$\text{logit}[P(Y \leq j|x)] = \alpha_j + \beta_1 x_1 + \beta_2 x_2.$$

[Table 8.6](#) shows output. With $J = 3$ response categories, the model has two $\{\alpha_j\}$ intercepts. Usually, these are not of interest except for computing response probabilities. The parameter estimates yield estimated logits and hence estimates of $P(Y \leq j)$, $P(Y > j)$, or $P(Y = j)$. We illustrate for white subjects ($x_2 = 0$) at the mean number of traumatic events score of $x_1 = 1.536$. Since $\hat{\alpha}_1 = -0.518$, the estimated probability of response *very happy* is

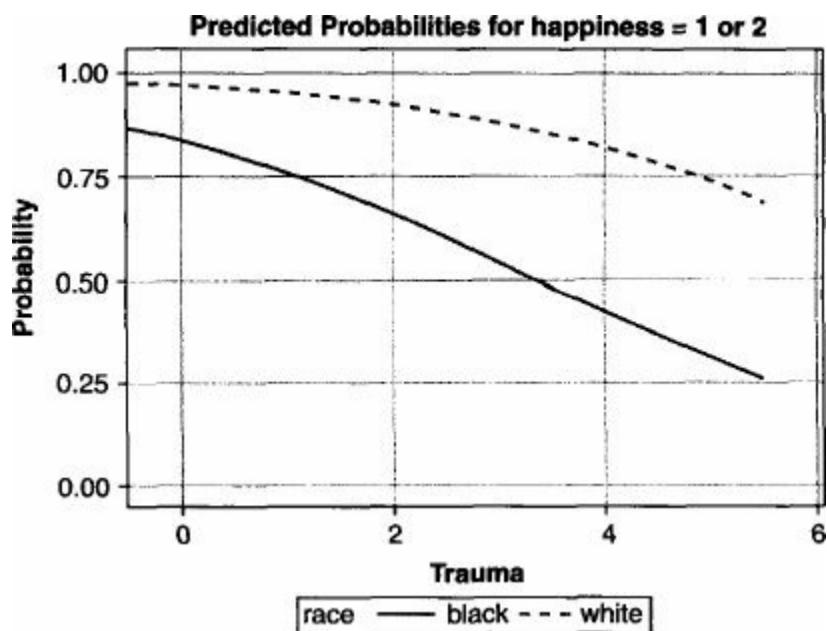
Table 8.6 Software Output (Based on SAS) for Fitting Cumulative Logit Model to Data on Happiness

Score Test for the Proportional Odds Assumption						
Chi-Square	DF	Pr > ChiSq				
0.8668	2	0.6483				
Parameter	Estimate	Std Error	Like. Ratio	95% Conf Limits	Chi-Square	Pr > ChiSq
Intercept1	-0.5181	0.3382	-1.2020	0.1392	2.35	0.1255
Intercept2	3.4006	0.5648	2.3779	4.6266	36.25	<.0001
traumatic	-0.4056	0.1809	-0.7729	-0.0520	5.03	0.0249
race	-2.0361	0.6911	-3.4287	-0.7156	8.68	0.0032

$$\hat{P}(Y = 1) = \hat{P}(Y \leq 1) = \frac{\exp[-0.518 - 0.406(1.536)]}{1 + \exp[-0.518 - 0.406(1.536)]} = 0.24.$$

[Figure 8.6](#) plots $\hat{P}(Y \leq 2)$ as a function of the number of traumatic events, at the two levels of race. An alternative way to portray the model is to plot the parallel straight lines for the fit in terms of the underlying latent variable.

[Figure 8.6](#) Estimated values of $P(Y \leq 2)$ by x_1 = number of traumatic events and x_2 = race.



The effect estimates $\hat{\beta}_1 = -0.406$ and $\hat{\beta}_2 = -2.036$ suggest that the cumulative probability starting at the very happy end of the happiness scale decreases as the traumatic events score increases and is lower for blacks than for whites. For example, given the traumatic events score, for whites the estimated odds of reporting being very happy were $e^{2.036} = 7.7$ times the estimated odds for blacks. This estimate is imprecise, because relatively few observations were in the black category. The 95% profile likelihood confidence interval for $-\hat{\beta}_2$ is $(0.72, 3.43)$, corresponding to $(2.05, 30.84)$ for the odds ratio effect. The SE values reported are based on the expected information from Fisher scoring. Using observed information (from Newton-Raphson), $\hat{\beta}_1$ and $\hat{\beta}_2$ have SE values of 0.183 and 0.686 instead of 0.181 and 0.691.

Descriptions of effects can compare cumulative probabilities rather than use odds ratios. These can make it easier to conceptualize the sizes of effects. We describe effects of quantitative variables by comparing probabilities at their extreme values or at their quartiles. We describe effects of qualitative variables by comparing probabilities for different categories. We fix values of quantitative variables by setting them at their mean or median. For qualitative variables we fix the category, unless there are several, in which case we can set each at their indicator means.

We illustrate again with $P(Y = 1)$, the *very happy* outcome. First, we describe the race effect. At the mean number of traumatic events of 1.536, $P(Y = 1) = 0.04$ for blacks (i.e., $x_2 = 1$) and 0.24 for whites ($x_2 = 0$). Next, we describe the number of traumatic events effect. The minimum and maximum values were 0 and 5. For blacks, $P(Y = 1)$ changes from 0.07 to 0.01 between these values; for whites, it changes from 0.37 to 0.07. (Note that comparing 0.07 to 0.37 at the minimum and 0.01 to 0.07 at the maximum provides further information about the race effect.) The sample effect is substantial for both predictors. However, these summaries are highly tentative and have large standard errors, because the black sample had only 13 observations, of whom none reported more than 3 traumatic events.

8.2.5 Checking the Proportional Odds Assumption

Models in this section used the proportional odds assumption of the same effects for different cumulative logits. An advantage is that effects are simple to summarize, requiring only a single parameter for each predictor. The models generalize to include separate effects, replacing β in (8.5) by β_j . This implies nonparallelism of curves for different logits. However, curves for different cumulative probabilities may then cross for some x values. Such models then violate the proper order among the cumulative probabilities (Exercise 8.37).

Even if such a model fits better over the observed range of x , for reasons of parsimony the simpler model might be preferable. One case is when effects $\{\beta_j\}$ with different logits are not substantially different in practical terms. Then the significance in a test of proportional odds may reflect primarily a large value of n . Even with smaller n , although effect estimators using the simpler model are biased, they may have smaller MSE than estimators from a model having many more parameters. So even if a test of proportional odds has a small P -value, don't discard this model automatically.

The output¹ in Table 8.6 also presents a score test of the proportional odds property. This tests whether the effects are the same for each cumulative logit against the alternative of separate effects. It compares the model with one parameter for x_1 and one for x_2 to the more complex model with two parameters for each, allowing different effects for $\text{logit}[P(Y \leq 1)]$ and $\text{logit}[P(Y \leq 2)]$. Here, the score statistic equals 0.87. It has $df = 2$, since the more complex model has two additional parameters. The more complex model does not fit significantly better ($P = 0.65$).

When this score test has a small P -value, it's helpful to check whether the violation of the proportional odds property is substantively important, by comparing estimates obtained from separate logistic fits to the binary collapsings of the response. For these data, consider the effect of the number of traumatic events. The model with binary response categories (very happy, pretty happy or not too happy) has $\beta_1 = -0.339$ ($SE = 0.213$), whereas the model with binary categories (very happy or pretty happy, not too happy) has $\beta_1 = -0.487$ ($SE = 0.276$). The effect has the same direction and a similar magnitude in each case, and it is sensible to use the simpler proportional odds structure. There is less information in the data about the race effect. We obtain $\beta_2 = -1.846$ ($SE = 0.763$) for the second collapsing but $\beta_2 = -\infty$ for the first collapsing because there were no observations for blacks in the very happy category and there is quasi-complete separation for that logit.

If a proportional odds model fits poorly in terms of practical as well as statistical significance, alternative strategies exist. These include (1) adding additional terms, such as interactions, to the linear predictor; (2) trying a link function for which the response curve is nonsymmetric (e.g., complementary log-log); (3) using an alternative ordinal model for which the more complex non-proportional-odds form is also valid; (4) adding dispersion parameters; (5) permitting separate effects for each logit for some but not all predictors (i.e., *partial proportional odds*); and (6) fitting baseline-category logit models and using the ordinality in an informal way in interpreting the associations.

For approach (1), more complex cumulative logit models are formulated as in ordinary logistic regression. For the example on modeling happiness, permitting interaction yields a model with ML fit

$$\text{logit}[\hat{P}(Y \leq j|x)] = \hat{\alpha}_j - 0.469x_1 - 3.057x_2 + 0.608(x_1x_2),$$

where the coefficient of x_1x_2 has $SE = 0.601$. The estimated effect of the number of traumatic events on the cumulative logit is -0.469 for whites and $(-0.469 + 0.608) = 0.139$ for blacks. The impact of the number of traumatic events may be quite different (and possibly nonexistent) for blacks, but recall that the black sample had only 13 observations, and here the difference in effects is not significant.

In the next section we generalize the cumulative logit model to permit extension (2) of alternative link functions. In Sections 8.3.4 and 8.3.6 we introduce models that satisfy option (3). Section 8.3.8 and Note 8.8 discuss extension (4). For approach (5), see Peterson and Harrell (1990), Stokes et al.

(2012), and criticism by Cox (1995). Agresti (2010, Chap. 3-5) discussed further these alternative strategies.

8.3 ORDINAL RESPONSES: ALTERNATIVE MODELS

Cumulative logit models use the logit link. As in binary GLMs, other link functions are possible. In this section we introduce models having alternative link functions either for cumulative probabilities or other response probabilities.

8.3.1 Cumulative Link Models

Let G^{-1} denote a link function that is the inverse of the continuous cdf G (recall Section 4.2.5). The *cumulative link* model

$$(8.7) \quad G^{-1}[P(Y \leq j | \mathbf{x})] = \alpha_j + \boldsymbol{\beta}^T \mathbf{x}$$

links the cumulative probabilities to the linear predictor. The logit link function $G^{-1}(u) = \log[u/(1 - u)]$ is the inverse of the standard logistic cdf.

As in the cumulative logit model with proportional odds form (8.5), effects of \mathbf{x} in (8.7) are the same for each cumulative probability. In Section 8.2.3 we showed that this assumption holds whenever a latent variable Y^* satisfies a linear regression model with standard cdf G for the error term. Model (8.7) results from discrete measurement of Y^* from a location-parameter family having cdf $G(y^* - \boldsymbol{\beta}^T \mathbf{x})$. The parameters $\{\alpha_j\}$ are category cutpoints (or “thresholds”) on a standardized version of the latent scale. Thus, we can regard cumulative link models as regression models that use a linear predictor $\boldsymbol{\beta}^T \mathbf{x}$ to describe effects of explanatory variables on crude ordinal measurement of Y^* . Using $-\boldsymbol{\beta}$ rather than $+\boldsymbol{\beta}$ in the linear predictor merely results in a change of sign of $\hat{\boldsymbol{\beta}}$.

8.3.2 Cumulative Probit and Log-Log Models

The *cumulative probit model* is the cumulative link model using the standard normal cdf Φ for G . This generalizes the binary probit model (Section 7.1) to ordinal responses. It is appropriate when the conditional distribution for the latent variable Y^* is normal. Parameters in probit models refer to effects on $E(Y^*)$. For instance, consider the model $\Phi^{-1}[P(Y \leq j)] = \alpha_j - \beta x$. From Section 8.2.3, since $Y^* = \beta x + \epsilon$ where $\epsilon \sim N(0,1)$ has cdf Φ , a 1-unit increase in x corresponds to a β increase in $E(Y^*)$. When ϵ need not be in standard form with $\sigma = 1$, a 1-unit increase in x corresponds to a β standard deviation increase in $E(Y^*)$.

Cumulative probit models provide fits similar to cumulative logit models. They have smaller estimates and standard errors because the standard normal distribution has standard deviation 1.0 compared with 1.81 for the standard logistic.

An underlying extreme value distribution for Y^* implies the model

$$\log\{-\log[1 - P(Y \leq j|x)]\} = \alpha_j + \boldsymbol{\beta}^T \mathbf{x}.$$

In Section 7.1 we introduced this *complementary log-log link* for binary data. The ordinal model using this link is sometimes called a *proportional hazards* model since it results from a generalization of the proportional hazards model for survival data to handle grouped survival times (McCullagh 1980, Note 8.6). It has the property

$$P(Y > j|x_1) = [P(Y > j|x_2)]^{\exp[\boldsymbol{\beta}^T(\mathbf{x}_1 - \mathbf{x}_2)]}.$$

With this link, $P(Y \leq j)$ approaches 1.0 at a faster rate than it approaches 0.0. The related *log-log link* $\log\{-\log[P(Y \leq j)]\}$ is appropriate when the complementary log-log link holds for the categories listed in reverse order.

McCullagh (1980) and Thompson and Baker (1981) treated cumulative link models as multivariate GLMs. McCullagh presented a Fisher scoring algorithm for ML estimation. He showed that sufficiently large n guarantees a unique maximum of the likelihood. Burridge (1981) and Pratt (1981) showed that the log likelihood is concave for many cumulative link models, including the logit, probit, and complementary log-log. Iterative algorithms usually converge rapidly to the ML estimates.

8.3.3 Example: Happiness Revisited with Cumulative Probits

In Section 8.2.4 we modeled $Y = \text{happiness}$ in terms of $x_1 = \text{total number of traumatic events}$ that happened to the respondent and his/her relatives in the last year, and $x_2 = \text{race}$. The cumulative logit model gave the fit

$$\text{logit}[\hat{P}(Y \leq j)] = \hat{\alpha}_j - 0.406x_1 - 2.036x_2.$$

The corresponding cumulative probit model has fit

$$\Phi^{-1}[\hat{P}(Y \leq j)] = \hat{\alpha}_j - 0.221x_1 - 1.157x_2,$$

with $SE = 0.098$ for $\hat{\beta}_1 = -0.221$ and $SE = 0.382$ for $\hat{\beta}_2 = -1.157$. The nature of the effects and the substantive significance is the same for the two models.

We can interpret parameter estimates in terms of the underlying latent variable model. For example, conditional on the number of traumatic events, the latent distribution on happiness is estimated to have location for whites that is 1.157 standard deviations in the more happy direction compared with that for blacks.

8.3.4 Adjacent-Categories Logit Models

Models for ordinal responses need not use cumulative probabilities. For the logit link, for example, ordinal logits can use pairs of adjacent response probabilities. The *adjacent-categories logits* are

$$(8.8) \quad \text{logit}[P(Y = j|Y = j \text{ or } j + 1)] = \log \frac{\pi_j}{\pi_{j+1}}, \quad j = 1, \dots, J - 1.$$

These logits are a basic set equivalent to the baseline-category logits. The connections are

$$(8.9) \quad \log \frac{\pi_j}{\pi_J} = \log \frac{\pi_j}{\pi_{j+1}} + \log \frac{\pi_{j+1}}{\pi_{j+2}} + \dots + \log \frac{\pi_{J-1}}{\pi_J}$$

and

$$\log \frac{\pi_j}{\pi_{j+1}} = \log \frac{\pi_j}{\pi_J} - \log \frac{\pi_{j+1}}{\pi_J}, \quad j = 1, \dots, J - 1.$$

Either set determines logits for all $\binom{J}{2}$ pairs of response categories.

Models using adjacent-categories logits can be expressed as baseline-category logit models. For instance, consider the adjacent-categories logit model of proportional odds form,

$$(8.10) \quad \log \frac{\pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x})} = \alpha_j + \boldsymbol{\beta}^T \mathbf{x}, \quad j = 1, \dots, J - 1,$$

with common effect $\boldsymbol{\beta}$. From adding $(J - j)$ terms as in (8.9), the equivalent baseline-category logit model is

$$\begin{aligned} \log \frac{\pi_j(\mathbf{x})}{\pi_J(\mathbf{x})} &= \sum_{k=j}^{J-1} \alpha_k + \boldsymbol{\beta}^T (J - j) \mathbf{x}, \quad j = 1, \dots, J - 1 \\ &= \alpha_j^* + \boldsymbol{\beta}^T \mathbf{u}_j, \quad j = 1, \dots, J - 1 \end{aligned}$$

with $\mathbf{u}_j = (J - j)\mathbf{x}$. The adjacent-categories logit model corresponds to a baseline-category logit model with adjusted model matrix but also a single parameter for each predictor.

The construction of the adjacent-categories logits recognizes the ordering of Y categories. To benefit from this in model parsimony requires appropriate specification of the linear predictor. When an explanatory variable has similar effect for each logit, advantages accrue from using the proportional odds form (8.10) with a single parameter instead of $(J - 1)$ parameters describing that effect. This model fits well in similar situations as the cumulative logit model of proportional odds form. Your choice of model type may reflect whether you prefer effects to refer to individual response categories, as the adjacent-categories logits provide, or instead to groupings of categories using the entire scale or an underlying latent variable, which cumulative logits provide. Since effects in cumulative logit models refer to the entire scale, they are usually larger in magnitude. The ratio of estimate to standard error, however, is usually similar for the two model types.

An advantage of the cumulative logit model is the approximate invariance of effect estimates to the choice and number of response categories. An advantage of the adjacent-categories logit model is that the more general model with $\boldsymbol{\beta}$ replaced by $\boldsymbol{\beta}_j$ is a valid model (i.e., cumulative probabilities will not be out of order), namely, one that is exactly equivalent to an ordinary baseline-category logit model. Also, because of its equivalence with canonical-link (baseline-category logit) models, the model has reduced sufficient statistics and we can use conditional ML estimation for inference with small samples or many parameters. Finally, its effects can be estimated with case-control studies (Mukherjee and Liu 2008).

8.3.5 Example: Happiness Revisited

We return to the example in Sections 8.2.4 and 8.3.3 on modeling happiness in terms of x_1 = total number of traumatic events and x_2 = race. The adjacent-categories logit model of proportional odds form has ML fit

$$\log[\hat{P}(Y = j)/\hat{P}(Y = j + 1)] = \hat{\alpha}_j - 0.357x_1 - 1.842x_2.$$

Conditional on the number of traumatic events, the estimated odds of being very happy instead of pretty happy, and the estimated odds of being pretty happy instead of not too happy, are $e^{1.842} = 6.31$ times as high for whites as for blacks. By contrast, the cumulative logit model had $\hat{\beta}_1 = -0.406$ and $\hat{\beta}_2 = -2.036$. As expected, its estimates are somewhat larger in magnitude. They are not much different for these data, however, because 65 of the 97 observations fall in the middle of the three response categories (pretty happy).

For these data, the more general model having different effects for each adjacent-categories logit has estimate $-\infty$ for the effect of race for the first logit, because there is quasi-complete separation for that logit. The estimates for the effect of number of traumatic events are -0.299 for the first logit and -0.432 for the second logit, suggesting that it is adequate to use the more parsimonious model of proportional odds form with its common estimate of -0.357 .

8.3.6 Continuation-Ratio Logit Models

The *continuation-ratio* logits are defined as

$$(8.11) \quad \log \frac{\pi_j}{\pi_{j+1} + \dots + \pi_J}, \quad j = 1, \dots, J-1$$

or as

$$(8.12) \quad \log \frac{\pi_{j+1}}{\pi_1 + \dots + \pi_j}, \quad j = 1, \dots, J-1.$$

The continuation-ratio logit model form is useful when a sequential mechanism, such as survival through various age periods, determines the response outcome (e.g., Tutz 1991). Let $\omega_j = P(Y=j|Y \geq j)$. With explanatory variables,

$$(8.13) \quad \omega_j(\mathbf{x}) = \frac{\pi_j(\mathbf{x})}{\pi_j(\mathbf{x}) + \dots + \pi_J(\mathbf{x})}, \quad j = 1, \dots, J-1.$$

The continuation-ratio logits (8.11) are ordinary logits of these conditional probabilities: namely, $\log[\omega_j(\mathbf{x})/(1 - \omega_j(\mathbf{x}))]$.

At the i th setting \mathbf{x}_i of \mathbf{x} , let $\{y_{ij}, j = 1, \dots, J\}$ denote the response counts, with $n_i = \sum_j y_{ij}$. When $n_i = 1$, y_{ij} indicates whether the response is in category j , as in Section 8.1.4. Let $b(n, y; \omega)$ denote the binomial probability of y successes in n trials with parameter ω for each trial. From the representation of the multinomial probability of (y_{i1}, \dots, y_{iJ}) in the form $p(y_{i1})p(y_{i2}|y_{i1}) \dots p(y_{iJ}|y_{i1}, \dots, y_{i,J-1})$, it follows that the multinomial mass function has factorization

$$(8.14) \quad b[n_i, y_{i1}; \omega_1(\mathbf{x}_i)]b[n_i - y_{i1}, y_{i2}; \omega_2(\mathbf{x}_i)] \dots b[n_i - y_{i1} - \dots - y_{i,J-2}, y_{i,J-1}; \omega_{J-1}(\mathbf{x}_i)].$$

The full likelihood is the product of multinomial mass functions from the different \mathbf{x}_i values. Thus, the log likelihood is a sum of terms such that different ω_j enter into different terms. When parameters in the model specification for $\text{logit}(\omega_j)$ are distinct from those for $\text{logit}(\omega_k)$ whenever $j \neq k$, maximizing each term separately maximizes the full log likelihood. Thus, separate fitting of models for different continuation-ratio logits gives the same results as simultaneous fitting. The sum of the $J-1$ separate G^2 statistics provides an overall goodness-of-fit statistic pertaining to the simultaneous fitting of $J-1$ models. Because of factorization (8.14), separate fitting can use methods for binary logistic models. Similar remarks apply to continuation-ratio logits (8.12), although those logits and the subsequent analysis do not give equivalent results.

Sometimes, a simpler proportional odds form of the model is plausible in which effects are the same for each logit (McCullagh and Nelder 1989, p. 164; Tutz 1991). Because of the factorization (8.14), it is also possible to fit such a model simply by creating a data file of independent binomials. See Agresti (2010, Sec. 4.2).

8.3.7 Example: Developmental Toxicity Study with Pregnant Mice

[Table 8.7](#) comes from a developmental toxicity study. Such experiments with rodents test substances posing potential danger to developing fetuses. Diethylene glycol dimethyl ether (diEGdiME), one such substance, is an industrial solvent used in the manufacture of protective coatings such as lacquer and metal coatings. This study administered diEGdiME in distilled water to pregnant mice. Each mouse was exposed to one of five concentration levels for 10 days early in the pregnancy. The mice exposed to level 0 formed a control group. Two days later, the uterine contents of the pregnant mice were examined for defects. Each fetus has three possible outcomes (nonlive, malformation, normal). The outcomes are ordered, with nonlive the least desirable result. We use continuation-ratio logits to model (1) the probability π_1 of a nonlive fetus, and (2) the conditional probability $\pi_2/(\pi_2 + \pi_3)$ of a malformed fetus, given that the fetus was live.

Table 8.7 Outcomes for Pregnant Mice in Developmental Toxicity Study

Concentration (mg/kg per day)	Response		
	Nonlive	Malformation	Normal
0 (controls)	15	1	281
62.5	17	0	225
125	22	7	283
250	38	59	202
500	144	132	9

^aBased on results in C. J. Price et al., *Fundam. Appl. Toxicol.* 8: 115–126, 1987.

I thank Louise Ryan for showing me these data.

We fitted the continuation-ratio logit models

$$\log \frac{\pi_1(x_i)}{\pi_2(x_i) + \pi_3(x_i)} = \alpha_1 + \beta_1 x_i, \quad \log \frac{\pi_2(x_i)}{\pi_3(x_i)} = \alpha_2 + \beta_2 x_i,$$

using x_i scores {0, 62.5, 125, 250, 500} for concentration level. The ML estimates are $\hat{\beta}_1 = 0.0064$ ($SE = 0.0004$) and $\hat{\beta}_2 = 0.0174$ ($SE = 0.0012$). In each case, the less desirable outcome is more likely as the concentration increases. For instance, given that a fetus was live, the estimated odds that it was malformed rather than normal multiplies by $\exp(1.74) = 5.7$ for every 100-unit increase in the concentration of diEGdiME. The likelihood-ratio fit statistics are $G^2 = 5.78$ for $j = 1$ and $G^2 = 6.06$ for $j = 2$, each based on $df = 3$. Their sum, $G^2 = 11.84$ (or similarly $X^2 = 9.76$), with $df = 6$, summarizes the fit.

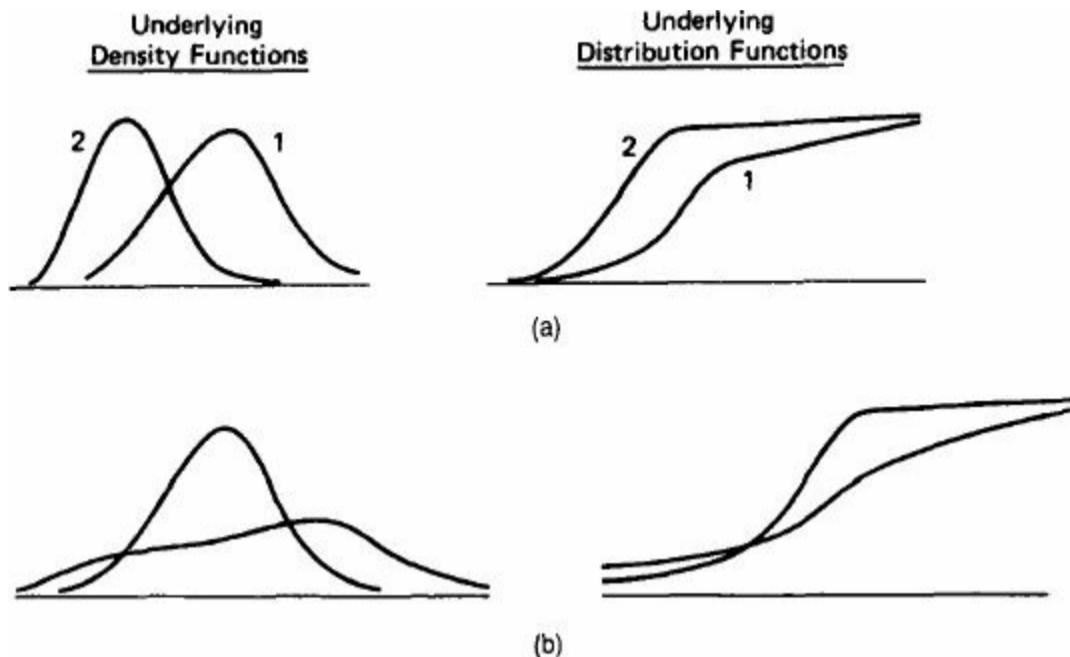
This analysis treats pregnancy outcomes for different fetuses as independent, identical observations. In fact, each pregnant mouse had a litter of fetuses, and statistical dependence may exist among different fetuses in the same litter. Different litters at a given concentration level may also have different response probabilities. Heterogeneity of various sorts among the litters (e.g., due to varying physical and/or genetic characteristics among different pregnant mice) would cause these probabilities to vary somewhat. Either statistical dependence or heterogeneous probabilities violates the binomial assumption and causes overdispersion. At a fixed concentration level, the number of fetuses in a litter that die may vary among pregnant mice more than if the counts were independent and identical binomial variates. The total G^2 shows some evidence of lack of fit ($P = 0.07$) but may reflect overdispersion caused by these factors rather than an inappropriate choice of response curve.

To account for overdispersion, we could adjust standard errors using the quasi-likelihood approach (Section 4.7). This multiplies standard errors by $\sqrt{X^2/df} = \sqrt{9.76/6} = 1.28$. For each logit, strong evidence remains that $\beta_j > 0$. In Chapters 13 and 14 we present other methods that account for the clustering of fetuses in litters.

8.3.8 Stochastic Ordering Location Effects Versus Dispersion Effects

For cumulative link models, settings of the explanatory variables are *stochastically ordered* on the response: For any pair \mathbf{x}_1 and \mathbf{x}_2 , either $P(Y \leq j|\mathbf{x}_1) \leq P(Y \leq j|\mathbf{x}_2)$ for all j or $P(Y \leq j|\mathbf{x}_1) \geq P(Y \leq j|\mathbf{x}_2)$ for all j . [Figure 8.7a](#) illustrates for underlying continuous density functions and cdf's at two settings of \mathbf{x} . Likewise, the adjacent-categories and continuation-ratio logit models with proportional odds structure imply stochastically ordered distributions for Y at different predictor values.

[Figure 8.7](#) (a) Distribution 1 stochastically higher than distribution 2; (b) distributions not stochastically ordered.



When this is violated and such models fit poorly, often it is because the dispersion also varies with \mathbf{x} . For instance, perhaps responses tend to concentrate around the same location but more dispersion occurs at \mathbf{x}_1 than at \mathbf{x}_2 . Then perhaps $P(Y \leq j|\mathbf{x}_1) > P(Y \leq j|\mathbf{x}_2)$ for small j but $P(Y \leq j|\mathbf{x}_1) < P(Y \leq j|\mathbf{x}_2)$ for large j . In other words, at \mathbf{x}_1 the responses concentrate more at the extreme categories than at \mathbf{x}_2 . [Figure 8.7b](#) illustrates for underlying continuous distributions.

Cumulative link models have been proposed that incorporate dispersion effects, but mainly for relatively simple cases such as with a single predictor that is a factor (Note 8.8). A simpler approach when a cumulative link model fits poorly is to fit the model separately for each cumulative probability to investigate the nature of the lack of fit or to use one of the other options mentioned at the end of Section 8.2.5.

8.3.9 Summarizing Predictive Power of Explanatory Variables

How can we summarize how well the response can be predicted using the fit of the chosen model? One approach estimates a measure such as the multiple correlation or R -squared for the regression model for an underlying latent response variable. McKelvey and Zavoina (1975) suggested this for the cumulative probit model.

Another index of predictive power generalizes the *concordance index* (Section 6.3.4). For all pairs of observations that have different response outcomes, it estimates the probability that the predictions and the outcomes are concordant, that is, that the observation with the larger y -value also has a stochastically higher set of estimated probabilities (and hence, for example, a higher mean for the estimated conditional distribution). The baseline value of no effect is 0.50. A value of 1.0 results when knowing which observation in an untied pair has the stochastically higher estimated distribution enables us to perfectly predict which one has the higher actual response. The higher the value of the concordance index, the better the predictive power.

Such measures are mainly useful for comparing different models. For example, for the happiness data analyzed with a proportional odds type of cumulative logit model in Section 8.2.4, the concordance index is 0.688 for the main-effects model and 0.689 when an interaction term is added. So, the more complex model is not much more useful for predictions, regardless of whether its extra term is statistically significant.

Keep in mind that predictive power is distinct from goodness of fit. A model may fit a particular data set well even if the predictive power the model provides is small. For other approaches to summarizing predictive power, see Agresti (2010, Sec. 3.4.6).

8.4 TESTING CONDITIONAL INDEPENDENCE IN $I \times J \times K$ TABLES

A common statistical analysis in many applications is studying whether an explanatory variable X has an effect on a response variable Y after we adjust for one or more other relevant factors. In Section 6.4 we considered this for binary Y and X using logistic models and the Cochran-Mantel-Haenszel (CMH) test of conditional independence for $2 \times 2 \times K$ tables. This section presents related tests with multicategory variables, in the context of $I \times J \times K$ tables. Likelihood-ratio tests compare the fit of a model specifying XY conditional independence with a model permitting X to have an effect. Generalizations of the CMH statistic are score statistics for certain models.

8.4.1 Testing Conditional Independence Using Multinomial Models

Denote a control factor by Z . Treating Z as nominal scale, we discuss four cases that treat (Y, X) as (nominal, nominal), (nominal, ordinal), (ordinal, nominal), (ordinal, ordinal). When Y is nominal, the baseline-category logit model of XY conditional independence is

$$(8.15) \quad \log \frac{P(Y = j|X = i, Z = k)}{P(Y = J|X = i, Z = k)} = \alpha_{jk}.$$

That is, each logit does not depend on the category of X . For ordinal Y we use cumulative logit models, but other ordinal models yield analogous tests. Then, XY conditional independence is equivalent to the model

$$\text{logit}[P(Y \leq j|X = i, Z = k)] = \alpha_{jk},$$

with $\alpha_{1k} < \alpha_{2k} < \dots < \alpha_{J-1,k}$ for each k . When the XY association is similar in the partial tables, the power of a test benefits from basing a test statistic on a model of homogeneous association.

1. Y nominal, \times nominal. An alternative to XY conditional independence that treats X as a factor is

$$(8.16) \quad \log \frac{P(Y = j|X = i, Z = k)}{P(Y = J|X = i, Z = k)} = \alpha_{jk} + \beta_{ij}$$

with constraint such as $\beta_{Ij} = 0$ for each j . For each outcome category, j , X and Z have additive effects of form $\alpha_k + \beta_i$. Conditional independence is $H_0: \beta_{1j} = \dots = \beta_{Jj}$ for $j = 1, \dots, J - 1$. Large-sample chi-squared tests have $\text{df} = (I - 1)(J - 1)$.

2. Y nominal, \times ordinal. Let $\{x_i\}$ be ordered scores. A test that is sensitive to the same linear trend alternatives in each partial table compares the conditional independence model to

$$\log \frac{P(Y = j|X = i, Z = k)}{P(Y = J|X = i, Z = k)} = \alpha_{jk} + \beta_j x_i.$$

Conditional independence is $H_0: \beta_1 = \dots = \beta_{J-1} = 0$. Large-sample chi-squared tests have $\text{df} = J - 1$.

3. Y ordinal, \times nominal. An alternative to XY conditional independence that treats X as a factor is

$$\text{logit}[P(Y \leq j|X = i, Z = k)] = \alpha_{jk} + \beta_i,$$

with a constraint such as $\beta_I = 0$. A simpler model that also has proportional odds structure for the effects of Z has linear predictor $\alpha_j + \beta_k^Z + \beta_i$. For either model, XY conditional independence is $H_0: \beta_1 = \dots = \beta_I$. Large-sample chi-squared tests have $\text{df} = I - 1$.

4. Y ordinal, X ordinal. For ordered scores $\{x_i\}$, the model

$$(8.17) \quad \text{logit}[P(Y \leq j|X = i, Z = k)] = \alpha_{jk} + \beta x_i$$

has the same linear trend for the X effect in each partial table. A simpler model that also has proportional odds structure for the effects of Z has linear predictor $\alpha_j + \beta_k^Z + \beta_i$. For either model, XY conditional independence is $H_0: \beta = 0$. Large-sample chi-squared tests have $\text{df} = 1$.

[Table 8.8](#) summarizes the four tests. They work well when the model describes a major component of the departure from conditional independence. This does not mean that we must test the fit of the model in order to use the test (see the remarks at the end of Section 6.4.2).

Table 8.8 Summary of Models for Testing Conditional Independence^a

$Y-X$	Model	Conditional Independence	df
Ordinal-ordinal	$\text{logit}[P(Y \leq j)] = \alpha_{jk} + \beta x_i$	$\beta = 0$	1
Ordinal-nominal	$\text{logit}[P(Y \leq j)] = \alpha_{jk} + \beta_i$	$\beta_1 = \dots = \beta_I$	$I - 1$
Nominal-ordinal	$\log \left[\frac{P(Y = j)}{P(Y = J)} \right] = \alpha_{jk} + \beta_j x_i$	$\beta_1 = \dots = \beta_{J-1} = 0$	$J - 1$
Nominal-nominal	$\log \left[\frac{P(Y = j)}{P(Y = J)} \right] = \alpha_{jk} + \beta_{ij}$	All $\beta_{ij} = 0$	$(I - 1)(J - 1)$

^aThe first two cases can also use $\alpha_j + \beta_k^Z$ in place of α_{jk} .

Occasionally, the association may change dramatically across the K partial tables. When Z is ordinal, an alternative by which a log odds ratio changes linearly across levels of Z is sometimes of use. For instance, when $Z = \text{age of subject}$, the association between a risk factor X (e.g., level of smoking) and a response Y (e.g., severity of heart disease) may tend to increase with Z . When Z is nominal, the conditional independence models can be compared with a more general alternative having separate effect parameters at each level of Z . Allowing effects to vary across levels of Z , however, results in the test df being multiplied by K , which handicaps power.

8.4.2 Example: Homosexual Marriage and Religious Fundamentalism

In 2008 the General Social Survey asked whether homosexuals should have the right to marry. One variable with which we'd expect responses to be associated is the fundamentalism/liberalism of a subject's religious beliefs. A subject's attained education is likely associated with both these variables, so is there an association when we condition on education? [Table 8.9](#) shows the relationship between opinion about homosexual marriage (Y) and religious beliefs (X), stratified by Z = attained education, for subjects of age 18–25.

Table 8.9 Opinion About Homosexual Marriage by Religious Beliefs, at Two Education Levels

Education	Religion	Homosexuals Should Be Able to Marry		
		Agree	Neutral	Disagree
High school or less	Fundamentalist	6	2	10
	Moderate	8	3	9
	Liberal	11	5	6
At least some college	Fundamentalist	4	2	11
	Moderate	21	3	5
	Liberal	22	4	1

Source: 2008 General Social Survey, subsample for ages 18–25.

[Table 8.10](#) summarizes the fit of several logistic models and shows the results of related likelihood-ratio tests of conditional independence. Each test compares a model to the model deleting the religious beliefs effect, conditioning on attained education. The models that treat opinion as ordinal use cumulative logits, with linear predictor $\alpha_j + \beta^Z_k + \beta x_i$ to treat X as an ordinal predictor using x_i scores (1, 2, 3) and linear predictor $\alpha_j + \beta^Z_k + \beta x_i$ to treat X as a nominal factor. The corresponding tests compare these to the model with linear predictor $\alpha_j + \beta^Z_k$. That model is not exactly equivalent to the conditional independence model [\(8.15\)](#), which is the last model listed in the table, with $G^2 = 26.85$ based on $df = 8$.

Table 8.10 Summary of Model-Based Likelihood-Ratio Tests of Conditional Independence for [Table 8.9](#)

Opinion	Religion	G^2 Fit	df	Test		
				Statistic	df	P-value
Ordinal	Ordinal	10.36	8	16.57	1	<0.0001
	Nominal	9.17	7	17.76	2	0.0001
	Not in model	26.93	9	—	—	—
Nominal	Ordinal	7.33	6	19.53	2	0.0001
	Nominal	6.58	4	20.27	4	0.0004
	Not in model	26.85	8	—	—	—

Testing conditional independence with the first cumulative logit model yields likelihood-ratio statistic $26.93 - 10.36 = 16.57$ with $df = 9 - 8 = 1$, strong evidence of an effect. Models that treat either or both variables as nominal also provide strong evidence, but not quite as strong. Focusing the test on a linear trend alternative yields a smaller P -value when that model describes reality reasonably well. However, we learn more from estimating model parameters than from these significance tests.

8.4.3 Generalized Cochran–Mantel–Haenszel Tests for $I \times J \times K$ Tables

The CMH statistic generalizes to multiple rows and columns. The tests treat X and Y symmetrically, so the three cases correspond to treating both as nominal, both as ordinal, or one of each. Conditional on row and column totals, each stratum has $(I - 1)(J - 1)$ nonredundant cell counts. Let

$$\mathbf{n}_k = (n_{11k}, n_{12k}, \dots, n_{1,J-1,k}, \dots, n_{I-1,J-1,k})^T.$$

Let $\boldsymbol{\mu}_k = E(\mathbf{n}_k)$ under H_0 : conditional independence, namely,

$$\boldsymbol{\mu}_k = (n_{1+k}n_{+1k}, n_{1+k}n_{+2k}, \dots, n_{I-1,+k}n_{+,J-1,k})^T / n_{++k}.$$

Let \mathbf{V}_k denote the null covariance matrix of \mathbf{n}_k , conditional on the margins, where

$$\text{cov}(n_{ijk}, n_{i'j'k}) = \frac{n_{i+k}(\delta_{ii'}n_{++k} - n_{i'+k})n_{+jk}(\delta_{jj'}n_{++k} - n_{+j'k})}{n_{++k}^2(n_{++k} - 1)}$$

with $\delta_{ab} = 1$ when $a = b$ and $\delta_{ab} = 0$ otherwise.

First, suppose the rows and columns are unordered. Let

$$\mathbf{n} = \sum_k \mathbf{n}_k, \quad \boldsymbol{\mu} = \sum_k \boldsymbol{\mu}_k, \quad \mathbf{V} = \sum_k \mathbf{V}_k.$$

The generalized CMH statistic for nominal X and Y is

$$(8.18) \quad \text{CMH} = (\mathbf{n} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{n} - \boldsymbol{\mu}).$$

Its large-sample chi-squared distribution has $\text{df} = (I - 1)(J - 1)$. The df value equals that for the statistics comparing logistic models (8.15) and (8.16). For $K = 1$ stratum with n observations, CMH = $[(n - 1)/n]X^2$, where X^2 is the Pearson statistic (3.10) for testing independence.

Next, suppose the rows and columns are both ordered. For ordered scores $\{u_i\}$ and $\{v_j\}$, evidence of a positive trend occurs if in each stratum $T_k = \sum_i \sum_j u_i v_j n_{ijk}$ exceeds its null expectation. Given the marginal totals, under conditional independence

$$E(T_k) = \left[\sum_i u_i n_{i+k} \right] \left[\sum_j v_j n_{+jk} \right] / n_{++k},$$

$$\text{var}(T_k) = \frac{1}{n_{++k} - 1} \left[\sum_i u_i^2 n_{i+k} - \frac{(\sum_i u_i n_{i+k})^2}{n_{++k}} \right]$$

$$\times \left[\sum_j v_j^2 n_{+jk} - \frac{(\sum_j v_j n_{+jk})^2}{n_{++k}} \right].$$

The statistic $[T_k - E(T_k)]/\sqrt{\text{var}(T_k)}$ equals the correlation between X and Y in stratum k multiplied by $\sqrt{n_{++k} - 1}$. To summarize across the K strata in a way that is sensitive to a correlation of common sign in each stratum, Mantel (1963) proposed

$$(8.19) \quad M^2 = \frac{\left\{ \sum_k \left[\sum_i \sum_j u_i v_j n_{ijk} - E \left(\sum_i \sum_j u_i v_j n_{ijk} \right) \right] \right\}^2}{\sum_k \text{var} \left(\sum_i \sum_j u_i v_j n_{ijk} \right)}.$$

This has a large-sample χ^2_1 null distribution, the same as for testing $H_0: \beta = 0$ in ordinal model (8.17). For $K = 1$, this is the M^2 correlation-based statistic (3.16).

Landis et al. (1978) presented a statistic that has (8.18) and (8.19) as special cases. Their statistic also can treat X as nominal and Y as ordinal, summarizing information about how I row means compare to their null expected values, with $\text{df} = I - 1$ (see Note 8.9).

8.4.4 Example: Homosexual Marriage Revisited

[Table 8.11](#) shows output for conducting generalized CMH tests with [Table 8.9](#). Statistics treating a variable as ordinal used scores (1, 2, 3) for opinion and for religious beliefs.

Table 8.11 Output (from SAS, PROC FREQ) for Generalized Cochran–Mantel–Haenszel Tests with Data from [Table 8.9](#)

Summary Statistics for opinion by religious fundamentalism Controlling for education				
Cochran–Mantel–Haenszel Statistics (Based on Table Scores)				
Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	16.83	<0.0001
2	Row Mean Scores Differ	2	17.94	0.0001
3	General Association	4	19.76	0.0006

The *general association* alternative treats X and Y as nominal and uses [\(8.18\)](#). It is sensitive to any association that is similar in each category of Z . The *nonzero correlation* alternative treats X and Y as ordinal and uses [\(8.19\)](#). It is sensitive to a similar linear trend in each category of Z . The *row mean scores differ* alternative treats rows as nominal and columns as ordinal. It is sensitive to variation among the I row mean scores on Y , when that variation is similar in each category of Z .

8.4.5 Related Score Tests for Multinomial Logit Models

The generalized CMH tests seem to be non-model-based alternatives to the tests of Section 8.4.1 using multinomial logit models. However, a close connection exists between them. For certain multinomial logit models, the generalized CMH tests are score tests of conditional independence.

The generalized CMH test (8.18) that treats X and Y as nominal is the score test that the $(I - 1)(J - 1)$ $\{\beta_{ij}\}$ parameters in model (8.16) equal 0. The generalized CMH test using M^2 that treats X and Y as ordinal is the score test of $\beta = 0$ in model (8.17). For the cumulative logit model, the equivalence has the same $\{x_i\}$ scores in the model as in M^2 , and the $\{v_j\}$ scores in M^2 are average rank scores. For the adjacent-categories logit model analog of (8.17), the $\{v_j\}$ scores in M^2 are any equally spaced scores.

With large samples in each stratum, the generalized CMH tests give similar results as likelihood-ratio tests comparing the relevant models. An advantage of the model-based approach is providing estimates of effects. An advantage of the generalized CMH tests is maintaining good performance under sparse asymptotics whereby K grows as n does. Also, they are valid under randomization arguments when there is not multinomial sampling from the population of interest but the multivariate hypergeometric distribution applies to each stratum under the null, such as for a volunteer sample of subjects randomly assigned to treatments in a clinical trial.

8.5 DISCRETE-CHOICE MODELS

Many applications of multinomial logit models relate to determining effects of explanatory variables on a subject's choice from a discrete set of options—for instance, transportation system to take to work (driving alone, carpooling, bus, subway, walk, bicycle), housing (buy house, buy condominium, rent), primary shopping location (downtown, mall, catalogs, Internet), or product brand. Models for response variables consisting of a discrete set of choices are called *discrete-choice models*.

8.5.1 Conditional Logits for Characteristics of the Choices

In many discrete-choice applications, an explanatory variable takes different values for different response choices. As predictors of choice of transportation system, the cost and the time to reach the destination take different values for each option. As a predictor of choice of product brand, the price varies according to the option. Explanatory variables of this type are *characteristics of the choices*. They differ from the usual ones, for which values remain constant across the choice set. Such variables, *characteristics of the chooser*, include demographic characteristics such as gender, race, and educational attainment.

McFadden (1974) proposed a discrete-choice model for explanatory variables that are characteristics of the choices. For subject i and response choice j , let $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijp})^T$ denote the values of the p explanatory variables, and let $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$. The model for the probability of selecting option j is

$$(8.20) \quad \pi_j(\mathbf{x}_i) = \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_{ij})}{\sum_h \exp(\boldsymbol{\beta}^T \mathbf{x}_{ih})}.$$

For each pair of choices a and b , this model has the logit form for conditional probabilities,

$$(8.21) \quad \log[\pi_a(\mathbf{x}_i)/\pi_b(\mathbf{x}_i)] = \boldsymbol{\beta}^T (\mathbf{x}_{ia} - \mathbf{x}_{ib}).$$

Conditional on the choice being a or b , a variable's influence depends on the distance between the subject's values of that variable for those choices. If the values are the same, the model asserts that the variable has no influence on the choice between a and b . Reflecting this property, McFadden originally referred to model (8.20) as a *conditional logit* model.

From (8.21), the odds of choosing a over b do not depend on the other alternatives in the choice set or on their values of the explanatory variables. Luce (1959) called this property *independence from irrelevant alternatives*. It is unrealistic in some applications. For instance, for travel options auto and red bus, suppose that 80% choose auto, corresponding to an odds of 4.0. Now suppose that the options are auto, red bus, and blue bus. According to (8.21), the odds are still 4.0 of choosing auto instead of red bus, but intuitively, we expect them to be about 8.0 (if about 10% choose each bus option). McFadden (1974) stated: “Application of the model should be limited to situations where the alternatives can plausibly be assumed to be distinct and weighed independently in the eyes of each decision-maker.”

McFadden’s model is actually a bit more general, permitting the choice set to vary among subjects. For instance, some subjects may not have the subway as an option for travel to work. In the denominator of (8.20), the sum is then taken over the choice set for subject i .

8.5.2 Multinomial Logit Model Expressed as Discrete-Choice Model

Discrete-choice models can also incorporate explanatory variables that are characteristics of the chooser. This may seem surprising, since formula (8.20) has a single parameter for each explanatory variable; that is, the parameter vector is the same for each pair of choices. However, multinomial logit model (8.2) has this discrete-choice form when we replace such an explanatory variable by J artificial variables. The j th is the product of the explanatory variable with a indicator variable that equals 1 when the response choice is j . For instance, for a single explanatory variable, let x_i denote its value for subject i . For $j = 1, \dots, J$, let δ_{jk} equal 1 when $k = j$ and 0 otherwise, and let

$$z_{ij} = (\delta_{j1}, \dots, \delta_{jJ}, \delta_{j1}x_i, \dots, \delta_{jJ}x_i)^T.$$

Let $\beta = (\alpha_1, \dots, \alpha_J, \beta_1, \dots, \beta_J)^T$. Then $\beta^T z_{ij} = \alpha_j + \beta_j x_i$, and (8.2) is (with $\alpha_J = \beta_J = 0$ for identifiability)

$$\begin{aligned}\pi_j(x_i) &= \frac{\exp(\alpha_j + \beta_j x_i)}{\exp(\alpha_1 + \beta_1 x_i) + \dots + \exp(\alpha_J + \beta_J x_i)} \\ &= \frac{\exp(\beta^T z_{ij})}{\exp(\beta^T z_{i1}) + \dots + \exp(\beta^T z_{iJ})}.\end{aligned}$$

This has the discrete-choice model form (8.20).

With this approach, discrete-choice models can contain characteristics of the chooser and of the choices. Thus, the model is very general. The ordinary multinomial logit model using baseline-category logits is a special case.

8.5.3 Example: Shopping Destination Choice

McFadden (1974) used discrete-choice models to describe how residents of Pittsburgh, Pennsylvania, chose a shopping destination. The five possible destinations were different city zones. One explanatory variable measured S = shopping opportunities, defined to be the retail employment in the zone as a percentage of total retail employment in the region. The other explanatory variable was P = price of the trip, defined from a separate analysis using auto in-vehicle time and auto operating cost.

The ML estimates of model parameters were -1.06 ($SE = 0.28$) for price of trip and 0.84 ($SE = 0.23$) for shopping opportunity. From [\(8.21\)](#),

$$\log(\hat{\pi}_a/\hat{\pi}_b) = -1.06(P_a - P_b) + 0.84(S_a - S_b).$$

Not surprisingly, a destination is relatively more attractive as the trip price decreases and as the shopping opportunity increases.

8.5.4 Multinomial Probit Discrete-Choice Models

Let U_{ij} denote the utility of alternative j for subject i . Suppose that

$$(8.22) \quad U_{ij} = \boldsymbol{\beta}^T \mathbf{x}_{ij} + \epsilon_{ij}$$

and the response choice is the value of j having maximum utility. McFadden (1974) showed that the assumption that $\{\epsilon_{ij}\}$ are independent and have the standard extreme value distribution is equivalent to discrete-choice model (8.20). (Recall Note 7.2 and Section 8.1.6.)

One way such a construction may be unrealistic is when the error terms partly represent unobserved covariates that are correlated with the response variable. Then, ϵ_{ia} and ϵ_{ib} are unlikely to be independent. However, this utility structure suggests alternative models, in particular, ones that do not have the property of independence from irrelevant alternatives. When we assume that $\boldsymbol{\epsilon} = (\epsilon_{i1}, \dots, \epsilon_{iJ})$ has a multivariate normal $N(\mathbf{0}, \Sigma)$ distribution, this utility model is a *multinomial probit model*, extending the model of Section 8.1.6. Model identifiability requires constraints on Σ , such as by taking $\text{var}(\epsilon_{i1}) = 1$ (Hausman and Wise 1978). Multinomial probit models are more complex computationally, requiring numerical integration or simulation to obtain the likelihood function.

8.5.5 Extensions: Nested Logit and Mixed Logit Models

In permitting correlated errors among response categories in a model for utilities, we could instead assume that ϵ has a multivariate form of extreme value distribution. This induces generalized logistic models. For example, McFadden considered applications in which the choice categories are partitioned into groups having a tree-like structure, with each group consisting of similar alternatives and having correlated error terms within groups. This is useful when the choices are naturally nested. An example is a person's choice of where to live: The person first chooses one of several communities to live in, and then within that community chooses a type of dwelling. Such a model for nested choices is called a *nested logit model*. Train (2009, pp. 77–88) gave an overview and multiple references.

Multinomial logit and probit discrete-choice models can be further generalized by treating certain effects as random rather than fixed, in the spirit of models considered later in this text in Chapters 12 and 13. A *mixed logit model* is one in which choice probabilities are obtained by integrating the logistic expression (8.20) for choice probabilities with respect to a distribution for certain model parameters. This allows heterogeneity among subjects in the size of effects. It is useful as a mechanism for inducing positive association among repeated responses with longitudinal data. Estimates of the parameters of the mixing distribution provide information about the average effects and the extent of the heterogeneity. Individual effects can also be predicted. For details, see McFadden (1974), Skrondal and Rabe-Hesketh (2004, Chap. 13), and Train (2009, Chap. 6).

8.5.6 Extensions: Discrete Choice with Ordered Categories

Sometimes the response categories have a natural ordering, such as the choice in renting a car among (subcompact, compact, midsize, large) size levels. Standard discrete-choice models do not account for such ordering. For multinomial logit models, the property of independence from irrelevant alternatives may then be especially unrealistic, as a particular response category is more similar to categories near it than categories further away.

Small (1987) proposed a model related to McFadden's multivariate extreme value model for utilities. In his model, the correlation between utility components for alternatives a and b is a nonincreasing function of $|a - b|$. Another approach uses a multinomial probit model with structure on the covariance matrix for ϵ that reflects the ordinality. For example, the correlation might have the autoregressive structure whereby $\text{corr}(\epsilon_{ia}, \epsilon_{ib}) = \rho^{|a-b|}$.

Beggs et al. (1981) considered an alternative type of ordered-alternatives problem in which subjects fully rank the outcome categories from best to worst. The categories need not themselves be ordered. They assumed the utility model (8.22), assuming *iid* extreme value errors. Let (r_{i1}, \dots, r_{iJ}) denote the ranking by subject i of the J choices, where r_{i1} is the response category given the highest ranking and r_{iJ} is the response category given the lowest ranking. Based on convenient properties of conditional distributions for extreme value distributions, they showed that that ranking vector for subject i has probability

$$P(U_{r_{i1}} > U_{r_{i2}} > \dots > U_{r_{iJ}}) = \prod_{h=1}^{J-1} \left[\exp(\boldsymbol{\beta}^T \mathbf{x}_{r_{ih}}) / \sum_{m=h}^J \exp(\boldsymbol{\beta}^T \mathbf{x}_{r_{im}}) \right].$$

Summing the logs of these terms over the n subjects yields the multinomial log likelihood. It can be maximized using Newton–Raphson methods. Beggs et al. (1981) applied the model to data in which various car types were ranked and the explanatory variables included car choice characteristics such as price, fuel cost, whether gas-powered or electric-powered, and subject socioeconomic family characteristics.

8.6 BAYESIAN MODELING OF MULTINOMIAL RESPONSES

The Bayesian approach for binary regression models extends to multinomial models. We focus here on Bayesian fitting of cumulative link models for ordinal responses and of multinomial (baseline-category) logit and probit models for nominal responses.

8.6.1 Bayesian Fitting of Cumulative Link Models

For an ordinal response Y , many models are special cases of the cumulative link model,

$$G^{-1}[P(Y \leq j|x)] = \alpha_j - \beta^T x.$$

From Section 8.2.3, this model is implied by a regression model for a latent variable having cdf G , such as logistic for the logit link. Prior distributions for the cutpoint parameters $\alpha = (\alpha_1, \dots, \alpha_{c-1})$ should take into account the ordering constraint

$$-\infty < \alpha_1 < \alpha_2 < \dots < \alpha_{c-1} < \infty.$$

In the cumulative probit case, the latent response for observation i is

$$Y_i^* = \beta^T x_i + \epsilon_i,$$

where $\{\epsilon_i\}$ are independent $N(0, 1)$. Albert and Chib (1993) presented a Bayesian analysis that utilizes the latent variable model and extends the analysis of Section 7.2.6 for binary responses. This model is simpler to handle than the cumulative logit model, because results apply from Bayesian inference for ordinary normal linear regression models, with a multivariate normal prior distribution for the regression parameters and independent normal latent variables $\mathbf{y}^* = (y_1^*, \dots, y_n^*)$. Implementation of MCMC methods is relatively simple because the Monte Carlo sampling is from normal distributions.

A Gibbs sampling scheme determines the posterior distribution by successively sampling from the density of (1) \mathbf{y}^* given \mathbf{y} , β , and α , (2) β given \mathbf{y} , \mathbf{y}^* , and α , and (3) α given \mathbf{y} , \mathbf{y}^* , and β . It uses the fact that if $y_i = j$, then y_i^* is between α_{j-1} and α_j . For example, given \mathbf{y} , β , and α , the conditional density function of y_i^* is normal with mean $\beta^T x_i$ and variance 1 but truncated between the two cutpoints corresponding to the value of y_i . Since \mathbf{y}^* determines \mathbf{y} , the conditional density of β given \mathbf{y} , \mathbf{y}^* , and α is proportional to the prior of β times the density of \mathbf{y}^* given β , which is normal since both components are normal. The conditional density function of α given \mathbf{y} , \mathbf{y}^* , and β is proportional to its truncated normal prior density but truncated to reflect that α_j must fall above all y_i^* such that $y_i = j$ but below all y_i^* such that $y_i = j + 1$.

Albert and Chib generalized the model to use link functions based on inverse cdf's for the t distribution. Since the logistic distribution relates closely to the t distribution with $df = 8$ (as described in Sections 4.2.5 and 7.2.6), this also provides a relatively simple way of fitting corresponding cumulative logit models. Alternatively, these days it is straightforward to use MCMC directly with the product of the chosen prior densities and the multinomial likelihood for the chosen model, regardless of the link function.

8.6.2 Example: Cannabis Use and Mother's Age

[Table 8.12](#) comes from a 21 –year follow-up study of mothers and their children who received antenatal care at a public hospital in Brisbane, Australia. At the age of 21, the children were asked “In the last month, how often did you use cannabis, marijuana, pot, etc.”? One explanatory variable was the mother’s age at entry to the study.

Table 8.12 Cannabis Use at 21 Years by Mother’s Age at Study Entry

Mother's Age	Cannabis Use at 21 Years				
	Never Use	Not Last Month	Once Last Month	Every Few Days	Every Day
<20 years	154	91	42	27	30
≥20 years	1078	567	261	157	111

Source: Hayatbakhsh et al. *Am. J. Drug & Alcohol Abuse*, 36: 350–356, 2010.

The deviance statistic for testing goodness of fit of the independence model in this table is $G^2 = 7.71$ ($df = 4, P = 0.10$). There is not much evidence of association, but this statistic ignores the ordinality of the response. Let’s consider the cumulative logit model of proportional odds form,

$$\text{logit}P(Y \leq j) = \alpha_j + \beta x,$$

with $x = 1$ for age ≥ 20 and $x = 0$ otherwise. The deviance is now 3.18 ($df = 3$). The ML estimate $\hat{\beta} = 0.230$ ($SE = 0.107$) and likelihood-ratio statistic of 4.53 ($df = 1, P = 0.033$) for testing $H_0: \beta = 0$ show considerable evidence that cannabis use tended to be lower when mother’s age was higher. The 95% profile likelihood confidence interval for β is (0.018, 0.441).

For a Bayesian analysis we recode x to take values 0.5 and -0.5 instead of 1 and 0, so the cumulative logits in each row have the same prior variability. For an analysis with uninformative priors, we used independent normal priors for the model parameters (appropriately truncated for $\{\alpha_j\}$ with means of 0 and standard deviations of 10. The posterior distribution of β , based on a Markov chain of length one million, then has a mean of 0.229 and a standard deviation of 0.108. The equal-tail 95% posterior interval for β is (0.018, 0.441), and the posterior $P(\beta < 0) = 0.017$. The sample size is large, so results are similar to those obtained with the frequentist approach. The Bayesian posterior $P(\beta < 0)$ is comparable to a frequentist one-sided P -value for $H_a: \beta > 0$.

8.6.3 Bayesian Fitting of Multinomial Logit and Probit Models

For nominal-scale responses, Albert and Chib (1993) presented Bayesian fitting of the multinomial probit model, using the connection with the latent utility model for maxima of normal random variates outlined in Section 8.1.6. We discuss this in terms of the discrete-choice form of the model outlined in Section 8.5.4, since standard models that do not have characteristics of the choices as explanatory variables are special cases. The underlying model for the utility for subject i making response j is

$$U_{ij} = \boldsymbol{\beta}^T \mathbf{x}_{ij} + \epsilon_{ij}$$

and the response choice is the value of j having maximum utility.

Let $\mathbf{U}_i = (U_{i1}, \dots, U_{iJ})$. When the errors are independent standard normal and we use a diffuse normal prior for $\boldsymbol{\beta}$, a Gibbs sampling scheme approximates the posterior distribution by successively sampling from the normal conditional densities of (1) $\boldsymbol{\beta}$ given \mathbf{y} , $\mathbf{U}_1, \dots, \mathbf{U}_N$, and (2) $\mathbf{U}_1, \dots, \mathbf{U}_N$ given \mathbf{y} and $\boldsymbol{\beta}$. In case (2) the distribution is truncated to reflect that if $y_i = j$ then component j of \mathbf{U}_i is its maximum. The model can also extend to let the utility components be correlated by introducing a parameter θ for the covariance matrix, such as a common correlation. A third step of the Gibbs sampling then includes sampling from its conditional density.

McCulloch et al. (2000) also dealt with multinomial probit models in terms of the underlying latent model. They noted the difficulty in placing priors on a covariance matrix that incorporate an identifiability constraint such as $\text{var}(U_{i1}) = 1$, and proposed priors that can account for that constraint. See also Imai and van Dyk (2005).

For baseline-category logit models, such routines do not connect with standard ones for normal variables, because the utility construction uses errors with extreme value distributions. However, it is not necessary to base computations on latent variable models or on the general discrete-choice version of the model, and with normal priors for the model parameters, software is widely available. With relatively diffuse priors, substantive results are usually similar to those with corresponding probit models.

Note, however, that if you place simple structure such as a common variance for the priors for $\beta_{1k}, \beta_{2k}, \dots, \beta_{J-1,k}$ in model (8.1), posterior results then depend somewhat on the choice of baseline category, because an effect relative to a pair of nonbaseline categories, $\beta_{jk} - \beta_{j'k}$, then has twice the prior variance. Alternatively, you can overparameterize by adding β_{Jk} to the model with the same prior but focus on the posterior differences for interpretation. The same remark applies to factors in such models, as results should ideally be invariant to the choice of a baseline category for indicators. One way to do this is to conduct the analysis in terms of corresponding Poisson loglinear models, introduced in the next chapter, which need not identify a baseline category for any categorical variable (Gelman et al. 2004, pp. 431–433). See www.stat.ufl.edu/~aa/cda/cda.html for details in terms of the following example. For examples of Bayesian uses of such models, see references cited in Note 8.12.

8.6.4 Example: Alligator Food Choice Revisited

For the alligator food choice data introduced in Section 8.1.2, we found that the probability of selecting a particular food choice was described well by a model with additive effects of size s and indicators contrasting lakes Hancock, Oklawaha, and Trafford with George. For the baseline choice of fish, the model is

Baseline Logit	Maximum Likelihood		Bayes, Prior $\sigma = 100$		Bayes, Prior $\sigma = 1$	
	$\hat{\beta}_{1j}$	SE	$\hat{\beta}_{1j}$	Std. Dev.	$\hat{\beta}_{1j}$	Std. Dev.
$\log(\pi_I/\pi_F)$	1.46	0.40	1.52	0.40	1.26	0.38
$\log(\pi_R/\pi_F)$	-0.35	0.58	-0.39	0.60	-0.55	0.48
$\log(\pi_B/\pi_F)$	-0.63	0.64	-0.68	0.67	-0.36	0.51
$\log(\pi_O/\pi_F)$	0.33	0.45	0.35	0.46	0.23	0.43

I, invertebrate; *R*, reptile; *B*, bird; *O*, other; *F*, fish.

As there was little prior information, especially about the lake effects, we fitted the model using diffuse independent normal prior distributions. [Table 8.13](#) shows posterior means and standard deviations for the size effect, when we parameterize in such a way that the 10 conditional log odds ratios relating size to pairs of food choice categories all have normal distributions with $\mu_0 = 0$ and $\sigma = 100$. Corresponding ML estimates and SE values are also shown. With such uninformative priors, results are quite similar. With either analysis, we conclude that the smaller alligators are relatively more likely to have invertebrates as their primary food choice.

Table 8.13 Estimated Size Effects and Standard Errors in Multinomial Logistic Model for Alligator Food Choice, Using Size and Lake as Predictors

$$\log(\pi_j/\pi_F) = \alpha_j + \beta_{1j}s + \beta_{2j}z_H + \beta_{3j}z_O + \beta_{4j}z_T, \quad j = 1, 2, 3, 4.$$

To compare with results from a highly informative Bayesian analysis, we used normal priors for the ten log odds ratios between size and pairs of food choices with each $\sigma = 1$. [Table 8.13](#) shows results. Having more prior information centered at 0 results in shrinkage of posterior estimates and standard deviations toward 0.

NOTES

Section 8.1: Nominal Responses: Baseline-Category Logit Models

8.1 BCL models: Baseline-category logit models were developed in Bock (1970), Haberman (1974a, pp. 352–373), Mantel (1966), Skrondal and Rabe-Hesketh (2003, 2004, Chap. 13), and Theil (1969, 1970). Lesaffre and Albert (1989) presented regression diagnostics. Amemiya (1981), Haberman (1982), and Theil (1970) presented R -squared measures. Baker (1994), Lang (1996), and Tsodikov and Chefo (2008) showed connections with Poisson models. Kosmidis and Firth (2011) used this connection in giving a penalized likelihood for bias reduction. Tutz and Schauberger (2012) proposed graphics for effects in multinomial response models.

Section 8.2: Ordinal Responses: Cumulative Logit Models

8.2 Cumulative logits: Early uses of cumulative logit models include Bock and Jones (1968), Simon (1974), Snell (1964), Walker and Duncan (1967), and Williams and Grizzle (1972). McCullagh (1980) popularized the proportional odds case. Later articles include Agresti and Lang (1993), Hastie and Tibshirani (1987), Peterson and Harrell (1990), and Tutz (1989). See also Note 12.2 and Sections 12.2.3 and 13.4.1.

8.3 Score test, power, efficiency: For $2 \times J$ tables and the model $\text{logit}[P(Y \leq j)] = \alpha_j + \beta x$, with x an indicator, McCullagh (1980) noted that the score test of $H_0: \beta = 0$ is equivalent to a discrete version of the Wilcoxon–Mann–Whitney test. Whitehead (1993) gave sample size formulas for this case. The sample size n_J needed for a certain power decreases as J increases: When response categories have equal probabilities, $n_J \asymp 0.75n_2/(1 - 1/J^2)$. The efficiency loss is major in collapsing to $J = 2$. See also Rabbee et al. (2003). Natarajan et al. (2012) extended the score test to complex sample survey data. Edwardes (1997) innovatively adapted the test by treating the cutpoints as random. Rice et al. (2012) discussed ways of dealing with variation in cutpoints.

8.4 ROC curve: As a way of evaluating diagnostic tests that have $J > 2$ ordered response categories rather than (positive, negative), an ROC curve can refer to the various possible cutoffs for defining a result to be positive. It plots sensitivity against 1 – specificity for the possible collapsings of the J categories to a (positive, negative) scale (Toledano and Gatsonis 1996).

Section 8.3: Ordinal Responses: Alternative Models

8.5 Probit, generalized links: Cumulative probit models were proposed by Aitchison and Silvey (1957) for the one-way layout setting and Gurland et al. (1960) and Bock and Jones (1968, Chap. 8) in a general regression setting. McKelvey and Zavoina (1975) presented the underlying latent normal model. Genter and Farewell (1985) introduced a generalized link function that permits comparison of fits provided by probit, complementary log–log, and other links. Adjacent-categories logit models and models equivalent to them were presented by Goodman (1979a, 1983), Haberman (1974b), and Simon (1974). Greene and Hensher (2010) presented other ordinal modeling strategies.

8.6 Hazard/survival: The ratio of a pdf to the complement of the cdf is the *hazard function* (Exercise 4.20). For discrete variables, this is the ratio found in continuation-ratio logits. The model $\text{log}[-\log(1 - \omega_j(x))] = \alpha_j + \beta^T x$ is a discrete-time version of the proportional hazards model (Allison 1982, Aranda-Ordaz 1983, Prentice and Gloeckler 1978, Thompson 1977). Läärä and Matthews (1985) showed this is equivalent to the model using the same link for cumulative probabilities.

8.7 OLS fitting: Assigning scores to ordered response categories and using ordinary least-squares regression modeling is not optimal, because the observations do not have constant

variance. Instead treating the response as multinomial, with categorical predictors Bhapkar (1968), Grizzle et al. (1969), and Williams and Grizzle (1972) used weighted least squares, and Haber (1985) and Lipsitz (1992) used ML. For large J , such models approximate a regression model for continuous Y . A structural difficulty is that the model can have predicted means outside the range of assigned scores. Also, “floor effects” and “ceiling effects” can occur when a latent response is categorized and a linear model is fitted to the observed response. See Agresti (2010, Sec. 1.3, 5.6) for details.

8.8 Dispersion effects: McCullagh (1980) generalized the cumulative link model to incorporate dispersion effects. With link function g , the model is

$$g[P(Y \leq j)] = \frac{\alpha_j - \beta^T x}{\exp(\gamma^T x)}.$$

The denominator contains scale parameters γ that describe how the dispersion depends on x . This model arises from a latent variable model in which the distribution of Y^* has shape reflected by g , such as normal for the probit link. The latent variable has $E(Y^*) = \beta^T x$ and standard deviation $\exp(\gamma^T x)$ that varies as x does. See also Agresti (2010, Sec. 5.4) and Cox (1995). Hamada and Wu (1990) and Nair (1987) presented alternatives models for detecting dispersion effects.

Section 8.4: Testing Conditional Independence in $I \times J \times K$ Tables

8.9 Generalized CMH: Birch (1965), Landis et al. (1978), Mantel (1963), and Mantel and Byar (1978) generalized the CMH statistic. Let the Kronecker product $B_k = u_k \otimes v_k$ denote a matrix of constants based on row scores u_k and column scores v_k for stratum k . The Landis et al. (1978) generalized statistic is

$$L^2 = \left[\sum_k B_k(n_k - \mu_k) \right]^T \left[\sum_k B_k V_k B_k^T \right]^{-1} \left[\sum_k B_k(n_k - \mu_k) \right].$$

When $u_k = (u_1, \dots, u_1)$ and $v_k = (v_1, \dots, v_J)$ for all strata, $L^2 = M^2$ in (8.19). When u_k is an $(I-1) \times I$ matrix $(I, -1)$, where I is an identity matrix of size $(I-1)$ and 1 denotes a column vector of $I-1$ ones, and v_k is the analogous matrix of size $(I-1) \times J$, L^2 simplifies to (8.18) with $df = (I-1)(J-1)$. With this u_k and $v_k = (v_1, \dots, v_J)$, L^2 sums over the strata information about how I row means compare to their null expected values, and it has $df = I-1$. Rank score versions are analogs for ordered categorical responses of stratum-adjusted Spearman correlation and Kruskal–Wallis tests. Kawaguchi et al. (2011) extended the Mantel–Haenszel odds ratio estimate to stratified Mann–Whitney estimators that utilize probability comparisons of two groups [related to Δ in (2.15)]. Landis et al. (2005) and Stokes et al. (2012) reviewed CMH methods. Koch et al. (1982) reviewed related methods.

8.10 Small-sample tests of conditional independence: To eliminate nuisance parameters, small-sample tests condition on row and column totals in each partial table. Section 7.3.5 showed this for $2 \times 2 \times K$ tables. When $I > 2$ and/or $J > 2$, the conditional distribution of cell counts in each stratum is the multivariate hypergeometric (Section 16.5.1), and this propagates an exact conditional distribution for the test statistic of interest, such as a generalized CMH statistic (Kim and Agresti 1997).

Section 8.5: Discrete-Choice Models

8.11 McFadden/Bradley–Terry/Luce: McFadden’s model relates to models proposed by Bradley and Terry (1952) (see Section 11.6) and Luce (1959). Train’s (2009) overview text includes many generalized models, and pages 45–50 discuss the independence from irrelevant alternatives assumption and references articles dealing with testing whether that property holds. One approach uses standard tests to compare it to a more complex nested logit model mentioned in Section 8.5.5.

Section 8.6: Bayesian Modeling of Multinomial Responses

8.12 Bayes multinomial: For other discussion of utilizing the connection with an underlying latent variable model, see Hoff (2009, Sec. 12.1) and Johnson and Albert (1999, Chap. 4). See also Congdon (2005, Chap. 7), and many references in Agresti (2010, Chap. 11). For comparing two ordinal categorical distributions, Altham (1969) provided a Bayesian estimate of the probability that one distribution is stochastically higher than the other. For Bayesian inference with baseline-category logit models, see Congdon (2005, Chap. 6), Daniels and Gatsonis (1997), Holmes and Held (2006), Leonard and Hsu (1994), and Sha et al. (2004).

EXERCISES

Applications

8.1 For [Table 8.14](#), let Y = belief in existence of heaven, x_1 = gender (1 = females, 0 = males), and x_2 = race (1 = blacks, 0 = whites). [Table 8.15](#) shows the fit of the model

Table 8.14 Data on Belief in Existence of Heaven for Exercise 8.1

Race	Gender	Belief in Heaven		
		Yes	Unsure	No
Black	Female	88	16	2
	Male	54	7	5
White	Female	397	141	24
	Male	235	189	39

Source: 2008 General Social Survey.

Table 8.15 Fit of Model for Belief in Heaven for Exercise 8.1

Belief Categories for Logit		
Parameter	Yes/No	Unsure/No
Intercept	1.785 (0.168)	1.554 (0.172)
Gender	1.044 (0.259)	0.254 (0.269)
Race	0.703 (0.411)	-0.106 (0.438)

$$\log(\pi_j/\pi_3) = \alpha_j + \beta_j^G x_1 + \beta_j^R x_2, \quad j = 1, 2,$$

with SE values in parentheses.

- a. Find the prediction equation for $\log(\pi_1\pi_2)$.
- b. Using the *yes* and *no* response categories, interpret the conditional gender effect using a 95% confidence interval for an odds ratio.
- c. Find $\hat{\pi}_1 = \hat{P}(Y = \text{yes})$ for white females.
- d. Without calculating estimated probabilities, explain why the intercept estimates indicate that for white males, $\hat{\pi}_1 > \hat{\pi}_2 > \hat{\pi}_3$. Use the intercept and gender estimates to show that the same ordering applies for black females.
- e. Without calculating estimated probabilities, explain why the estimates in the gender row indicate that $\hat{\pi}_1$ is higher for females than for males, for each race.
- f. For this fit, $G^2 = 0.69$. Explain why residual df = 2. Deleting the gender effect, $G^2 = 47.64$. Conduct a likelihood-ratio test of whether opinion is independent of gender, given race. Interpret.

8.2 A model fit predicting preference for U.S. President (Democrat, Republican, Independent) using x = annual income (in \$10,000) is $\log(\hat{\pi}_D/\hat{\pi}_I) = 3.3 - 0.2x$ and $\log(\hat{\pi}_R/\hat{\pi}_I) = 1.0 + 0.3x$.

- a. Find the prediction equation for $\log(\hat{\pi}_R/\hat{\pi}_D)$ and interpret the slope. For what range of x is $\hat{\pi}_R > \hat{\pi}_D$?
- b. Find the prediction equation for $\hat{\pi}_I$.
- c. Plot $\hat{\pi}_D$, $\hat{\pi}_I$, and $\hat{\pi}_R$ for x between 0 and 10, and interpret.

8.3 [Table 8.16](#) shows recent GSS data for the effect of gender and race on political party identification. Find a baseline-category logit model that fits well. Interpret estimated effects on the odds that party identification is Democrat instead of Republican.

Table 8.16 Data for Exercise 8.3 on Political Party ID

		Political Party Identification		
Gender	Race	Democrat	Republican	Independent
Male	White	132	176	127
	Black	42	6	12
Female	White	172	129	130
	Black	56	4	15

8.4 For 63 alligators caught in Lake George, Florida, [Table 8.17](#) classifies primary food choice as (fish, invertebrate, other) and shows length in meters. Alligators are called subadults if length < 1.83 meters (6 feet) and adults if length > 1.83 meters.

Table 8.17 Data for Exercise 8.4^a on Alligator Food Choice

Males				Females			
Length (m)	Choice	Length (m)	Choice	Length (m)	Choice	Length (m)	Choice
1.30	I	1.70	I	3.33	F	1.78	O
1.32	F	1.73	O	3.56	F	1.80	I
1.32	F	1.78	F	3.58	F	1.88	I
1.40	F	1.78	O	3.66	F	2.16	F
1.42	I	1.80	F	3.68	O	2.26	F
1.42	F	1.85	F	3.71	F	2.31	F
1.47	I	1.93	I	3.89	F	2.36	F
1.47	F	1.93	F	1.24	I	2.39	F
1.50	I	1.98	I	1.30	I	2.41	F
1.52	I	2.03	F	1.45	I	2.44	F
1.63	I	2.03	F	1.45	O	2.56	O
1.65	O	2.31	F	1.55	I	2.67	F
1.65	O	2.36	F	1.60	I	2.72	I
1.65	I	2.46	F	1.60	I	2.79	F
1.65	F	3.25	O	1.65	F	2.84	F
1.68	F	3.28	O	1.78	I		

^aF, fish; I, invertebrates; O, other.

- a.** Measuring length as (adult, subadult), find a model that adequately describes effects of gender and length on food choice. Interpret the effects. For adult females, find the estimated probabilities of the food choice categories.
- b.** Using only observations for which primary food choice was fish or invertebrate, find a model that adequately describes effects of gender and binary length. Compare parameter estimates and standard errors for this separate-fitting approach to those obtained with simultaneous fitting, including the other category.
- c.** Treating length as binary loses information. Adapt the model in part (a) to use the continuous length measurements. Interpret, explaining how the estimated outcome probabilities vary with length. Find the estimated length at which the invertebrate and other categories are equally likely.

8.5 Fit the multinomial probit model to the alligator food choice data in [Table 8.1](#) and at the text website, with size and lake as predictors. Compare estimates and SE values to those in [Table 8.4](#), and explain why they are larger for the multinomial logit model.

8.6 Fit the baseline-category logit model with main effects to the data in [Table 8.5](#). Describe the effect of the sample having no blacks in the very happy category.

8.7 For recent GSS data, the cumulative logit model [\(8.5\)](#) with Y = political ideology (very liberal, slightly liberal, moderate, slightly conservative, very conservative) and x = party affiliation (1 for the 428 Democrats and 0 for the 407 Republicans) has $\hat{\beta} = 0.975$ ($SE = 0.129$) and $\hat{\alpha}_1 = -2.469$. Interpret $\hat{\beta}$. Find the estimated probability of a very liberal response for each group.

8.8 [Table 8.18](#) is an expanded version of a data set analyzed in Section 9.4.2. The response categories are (1) not injured, (2) injured but not transported by emergency medical services, (3)

injured and transported by emergency medical services but not hospitalized, (4) injured and hospitalized but did not die, and (5) injured and died. [Table 8.19](#) shows output for a model of form [\(8.5\)](#).

[Table 8.18](#) Data for Exercise 8.8 on Degree of Injury in Auto Accident

Gender	Location	Seat Belt	Response on Injury Outcome				
			1	2	3	4	5
Female	Urban	No	7,287	175	720	91	10
		Yes	11,587	126	577	48	8
	Rural	No	3,246	73	710	159	31
		Yes	6,134	94	564	82	17
Male	Urban	No	10,381	136	566	96	14
		Yes	10,969	83	259	37	1
	Rural	No	6,123	141	710	188	45
		Yes	6,693	74	353	74	12

Source: Data courtesy of Cristanna Cook, Medical Care Development, Augusta, Maine.

[Table 8.19](#) Output for Exercise 8.8 on Auto Accident Injuries

Parameter		DF	Estimate	Std Error
Intercept1		1	3.3074	0.0351
Intercept2		1	3.4818	0.0355
Intercept3		1	5.3494	0.0470
Intercept4		1	7.2563	0.0914
gender	female	1	-0.5463	0.0272
gender	male	0	0.0000	0.0000
location	rural	1	-0.6988	0.0424
location	urban	0	0.0000	0.0000
seatbelt	no	1	-0.7602	0.0393
seatbelt	yes	0	0.0000	0.0000
location*seatbelt	rural no	1	-0.1244	0.0548
location*seatbelt	rural yes	0	0.0000	0.0000
location*seatbelt	urban no	0	0.0000	0.0000
location*seatbelt	urban yes	0	0.0000	0.0000

- a. Why are there four intercepts? Explain how they determine the estimated response distribution for males in urban areas wearing seat belts.
- b. Construct a confidence interval for the effect of gender, given seat-belt use and location. Interpret.
- c. Find the estimated cumulative odds ratio between the response and seat-belt use for those in rural locations and for those in urban locations, given gender. Based on this, explain how the effect of seat-belt use varies by region, and explain how to interpret the interaction estimate, -0.1244.

8.9 In a class project, University of Florida students Shahrzad Farshi and Marty Parks used GSS data to study the effect of several explanatory variables on liking for rap music, an ordinal variable with five categories (1 = greatest preference). They found a good fit with the model $\text{logit}[P(Y \leq j)] = \alpha_j - 1.06r - 0.58a$, where race r was coded 1 for white and 0 for black/other and age a has scores (1, 2, 3, 4) for four successive age categories. Interpret these effects with cumulative odds ratios.

8.10 [Table 8.20](#) refers to a clinical trial for the treatment of small-cell lung cancer. Patients were randomly assigned to two treatment groups. The sequential therapy administered the same combination of chemotherapeutic agents in each treatment cycle; the alternating therapy had three different combinations, alternating from cycle to cycle.

[Table 8.20](#) Data for Exercise 8.10 on Lung Cancer Clinical Trial

Therapy	Gender	Response to Chemotherapy			
		Progressive Disease	No Change	Partial Remission	Complete Remission
Sequential	Male	28	45	29	26
	Female	4	12	5	2
Alternating	Male	41	44	20	20
	Female	12	7	3	1

Source: W. Holtbrugge and M. Schumacher, *Appl. Statist.* **40**: 249–259, 1991.

- a. Fit a cumulative logit model with main effects for therapy and gender. Interpret effect estimates.
- b. For the therapy effect, compare $\hat{\beta}_1$ to the estimate obtained when the model is fitted to the binary response obtained by combining the first two response categories and combining the last two response categories. What property of the model does this reflect?
- c. For the collapsing in (b), compare $\hat{\beta}_1/SE$ to the ratio obtained for the uncollapsed response. (Usually, a disadvantage of collapsing ordinal responses is that the significance of effects diminishes.)
- d. Fit the model to the uncollapsed data that also contains an interaction term. Interpret. Does it fit better? Explain why it is equivalent to using the four gender–therapy combinations as levels of a single factor.

8.11 A study of factors affecting alcohol consumption measures the response variable with the scale (abstinence, a drink a day or less, more than one drink a day). For a comparison of two groups while adjusting for relevant covariates, the researchers hypothesize that the two groups will have about the same prevalence of abstinence, but that one group will have a considerably higher proportion who have more than one drink a day. Even though the response variable is ordinal, explain why a cumulative logit model with proportional odds structure may be inadequate for this study.

8.12 Refer to [Table 8.14](#). Treating belief in heaven as ordinal, fit and interpret a (a) cumulative logit model and (b) cumulative probit model. Compare results and state interpretations in each case.

8.13 For the cumulative probit model fitted to [Table 8.5](#), find the means and standard deviation for the two normal cdf's that provide the curves for $P(Y = 1)$ as a function of x_1 = number of traumatic events, at the two levels of x_2 = race. Interpret the effects.

8.14 For [Table 8.5](#), fit and interpret effects for a (a) cumulative link model with complementary log–log link and (b) continuation-ratio logit model.

8.15 Refer to Exercise 8.7. With adjacent-categories logit model [\(8.10\)](#), $\beta = 0.435$. Interpret using odds ratios for adjacent categories and for the (very liberal, very conservative) pair of categories.

8.16 For the developmental toxicity data in [Table 8.7](#), formulate and fit a continuation-ratio logit model with proportional odds structure. [Hint: Create a data file of independent binomials and then construct a model matrix that has the desired model structure.]

8.17 [Table 8.21](#) refers to a study that randomly assigned subjects to a control or treatment group. Daily during the study, treatment subjects ate cereal containing psyllium. The study analyzed the effect on LDL cholesterol.

Table 8.21 Data for Exercise 8.17 on Cholesterol and Cereal

Beginning	Ending LDL Cholesterol Level							
	Control				Treatment			
	≤ 3.4	3.4–4.1	4.1–4.9	> 4.9	3.4	3.4–4.1	4.1–4.9	> 4.9
≤ 3.4	18	8	0	0	21	4	2	0
3.4–4.1	16	30	13	2	17	25	6	0
4.1–4.9	0	14	28	7	11	35	36	6
> 4.9	0	2	15	22	1	5	14	12

Source: Data courtesy of Sallee Anderson, Kellogg Co.

- a. Model the ending cholesterol level as a function of treatment, using the beginning level as a covariate. Interpret the treatment effect.

- b. Repeat part (a), now treating the beginning level as qualitative. Compare results.

8.18 The book's website (www.stat.ufl.edu/~aa/cda/cda.html) has a $3 \times 4 \times 4$ table that cross-classifies dumping severity (Y) and operation (X) for four hospitals (H). The four operations refer to treatments for duodenal ulcer patients and have a natural ordering. Dumping severity describes a possible undesirable side effect of the operation. Its three categories are also ordered. [Table 8.22](#) shows results of generalized CMH tests. For each test, give a pair of models such that a likelihood-ratio test comparing those models would give similar results. Explain how one test can be much more significant than the others.

Table 8.22 Results for Dumping Severity Data of Exercise 8.18

Statistic	Summary Statistics for dumping by operate Controlling for hospital			
	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	6.3404	0.0118
2	Row Mean Scores Differ	3	6.5901	0.0862
3	General Association	6	10.5983	0.1016

8.19 A sample of subjects indicate their favorite among four Margarita pizzas characterized by 1 = (thin crust, normal cheese), 2 = (thin crust, extra cheese), 3 = (thick crust, normal cheese), 4 = (thick crust, extra cheese). For the characteristics of the choices x_1 = crust type (1 = thick, 0 = thin) and x_2 = cheese quantity (1 = extra, 0 = normal), the multinomial discrete choice model ([8.20](#)) has $\beta_1 = -0.40$ and $\beta_2 = 0.60$. For each pizza type, find the probability that it is the favorite.

8.20 Refer to the previous exercise. For a random sample of 20 pizza lovers, suppose 4 prefer choice 1, 8 prefer choice 2, 3 prefer choice 3, and 5 prefer choice 4. Fit the model and interpret the estimates.

8.21 Describe an application in which a discrete-choice model would be useful. Specify potential explanatory variables, and identify which are characteristics of the chooser and which are characteristics of the choices.

8.22 A cafe has four entrées: chicken, beef, fish, vegetarian. Specify a model of form ([8.20](#)) for the selection of an entrée using x = gender (1 = female, 0 = male) and u = cost of entrée, which is a characteristic of the choices. Interpret the model parameters.

8.23 For [Table 8.14](#) on belief in heaven, use Bayesian methods to fit the model of Exercise 8.1. Do this once with uninformative priors (say, $\sigma = 100$) and once with very informative priors (say, $\sigma = 1$). In each case, for the gender effect on the (yes/no) logit, report the posterior mean and standard deviation and the 95% posterior interval. Compare results between them and with the ML estimate, SE , and 95% confidence interval.

8.24 In the previous exercise, treat belief in heaven as ordinal and reanalyze with Bayesian methods. Compare results for the gender effect, and interpret.

8.25 Consider the baseline-category logit model of Section 8.6.4 for Bayesian modeling of alligator food choice in terms of size and lake. Try to replicate results in [Table 8.13](#) for $\sigma = 1$. (If your results differ much, for your parameterization the 10 conditional log odds ratios relating size to pairs of food choices may not all have prior $\sigma = 1$.)

8.26 Is political ideology associated with happiness? Conduct a Bayesian analysis for the data in [Table 3.7](#), using a model presented in this chapter. Present a posterior interval and posterior probability that addresses the question, and interpret results.

8.27 Analyze [Table 8.5](#) with two types of model studied in this chapter. Write a report summarizing results and advantages and disadvantages of each modeling strategy.

8.28 This book's website has a $4 \times 2 \times 3 \times 3$ table that cross-classifies a sample of residents of Copenhagen on type of housing (H), degree of contact with other residents (C), feeling of influence on apartment management (I), and satisfaction with housing conditions (S). Treating S as the response variable, analyze these data.

Theory and Methods

8.29 A multivariate generalization of the exponential dispersion family [\(4.17\)](#) is

$$f(\mathbf{y}_i; \boldsymbol{\theta}_i, \phi) = \exp\{[\mathbf{y}_i^T \boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)]/\phi + c(\mathbf{y}_i, \phi)\},$$

where $\boldsymbol{\theta}_i$ is the natural parameter. Show that a multinomial variate y_i for a single trial with parameters $\{\pi_j, j = 1, \dots, J-1\}$ is in the $(J-1)$ -parameter exponential family, with baseline-category logits as natural parameters.

8.30 Cell counts $\{y_{ij}\}$ in an $I \times J$ contingency table have a multinomial $(n; \{\pi_{ij}\})$ distribution. Show that $\{P(Y_{ij} = n_{ij})\}$ can be expressed as

$$\begin{aligned} d^n n! \prod_i \prod_j (n_{ij}!)^{-1} \exp \left[& \sum_{i=1}^{I-1} \sum_{j=1}^{J-1} n_{ij} \log(\alpha_{ij}) \right. \\ & \left. + \sum_{i=1}^{I-1} n_{i+} \log(\pi_{iJ}/\pi_{IJ}) + \sum_{j=1}^{J-1} n_{+j} \log(\pi_{Ij}/\pi_{IJ}) \right], \end{aligned}$$

where $\alpha_{ij} = \pi_{ij}\pi_{IJ}/\pi_{iJ}\pi_{Ij}$ and d is a constant independent of the data. Find an alternative expression using local odds ratios $\{\theta_{ij}\}$, by showing that

$$\sum_i \sum_j n_{ij} \log \alpha_{ij} = \sum_i \sum_j s_{ij} \log \theta_{ij}, \quad \text{where } s_{ij} = \sum_{a \leq i} \sum_{b \leq j} n_{ab}.$$

(Hence, models for such parameters have reduced sufficient statistics and relatively simple score statistics for testing effects.)

8.31 Consider the baseline-category logit model expressed as

$$\pi_j(\mathbf{x}) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j^T \mathbf{x})}{\sum_{h=1}^J \exp(\alpha_h + \boldsymbol{\beta}_h^T \mathbf{x})}.$$

Show that dividing numerator and denominator by $\exp(\alpha_J + \boldsymbol{\beta}_J^T \mathbf{x})$ yields new parameters $\alpha_j^* = \alpha_j - \alpha_J$ and $\boldsymbol{\beta}_j^* = \boldsymbol{\beta}_j - \boldsymbol{\beta}_J$ that satisfy $\alpha_J = 0$ and $\boldsymbol{\beta}_J = \mathbf{0}$. Thus, without loss of generality, we can take $\alpha_J = 0$ and $\boldsymbol{\beta}_J = \mathbf{0}$.

8.32 When there are $J = 3$ outcome categories, suppose that

$$\pi_j(x) = \exp(\alpha_j + \beta_j x)/[1 + \exp(\alpha_1 + \beta_1 x) + \exp(\alpha_2 + \beta_2 x)],$$

$j = 1, 2$. Show that $\pi_3(x)$ is **(a)** decreasing in x if $\beta_1 > 0$ and $\beta_2 > 0$, **(b)** increasing in x if $\beta_1 < 0$ and $\beta_2 < 0$, and **(c)** nonmonotone when β_1 and β_2 have different signs.

8.33 Refer to the log-likelihood function for the baseline-category logit model (Section 8.1.4). Denote the sufficient statistics by $\mathbf{n} \mathbf{p}_j = \sum_i y_{ij}$ and $S_{jk} = \sum_i x_{ik} y_{ij}, j = 1, \dots, J, k = 1, \dots, p$. Let $\mathbf{S} = (S_{11}, \dots, S_{1p}, \dots, S_{J1}, \dots, S_{Jp})^T$. Under the null hypothesis that explanatory variables have no effect, conditional on $\sum_i y_{ij}, j = 1, \dots, J$, show that

$$E(\mathbf{S}) = n(\mathbf{p} \otimes \mathbf{m}), \quad \text{var}(\mathbf{S}) = n(\mathbf{V} \otimes \boldsymbol{\Sigma}),$$

where $\mathbf{p} = (p_1, \dots, p_J)^T$, $\mathbf{m} = (\bar{x}_1, \dots, \bar{x}_p)^T$ with $\bar{x}_k = (\sum_i x_{ik})/n$, $\boldsymbol{\Sigma}$ has elements $s_{kv}^2 = [\sum_i (x_{ik} - \bar{x}_k)(x_{iv} - \bar{x}_v)]/(n-1)$, and \mathbf{V} has elements $v_{ii} = p_i(1-p_i)$ and $v_{ij} = -p_i p_j$ (Zelen 1991).

8.34 An alternative fitting approach for the baseline-category logit model (8.1) fits binary logistic models separately for the $J - 1$ pairings of responses. The estimates have larger SE than the ML estimates for simultaneous fitting of the $J - 1$ logits, but Begg and Gray (1984) showed that the efficiency loss is minor when the response category having highest prevalence is the baseline. Illustrate, by showing that the fit using categories I and F alone of the alligator data is $\log(\hat{\pi}_I/\hat{\pi}_F) = -1.69 + 1.66s - 1.78z_H + 1.05z_O + 1.22z_T$, with SE values (0.43, 0.62, 0.49, 0.52) for the effects. Compare with the first row of Table 8.4.

8.35 For explanatory variable k in a baseline-category logit model, suppose the model matrix constrains $\beta_{2k} = \dots = \beta_{Jk} = 0$, leaving β_{1k} unconstrained. Explain how β_{1k} then describes a contrast for that variable between outcome category 1 and the other categories combined. Explain how to generalize this to contrast one subset of the categories to the other categories.

8.36 Explain why the cumulative logit model of proportional odds form is not a special case of a baseline-category logit model.

8.37 Consider the cumulative logit model, $\text{logit}[P(Y \leq j)] = \alpha_j + \beta_j x$, not having proportional odds form.

- a. With continuous x taking values over the real line, show that the model is improper in that cumulative probabilities are misordered for a range of x values.
- b. When x is a binary indicator, explain why the model is proper but requires constraints on $(\alpha_j + \beta_j)$ (as well as the usual ordering constraint on $\{\alpha_j\}$) and is then equivalent to the saturated model.

8.38 For an $I \times J$ contingency table with ordinal Y and scores $\{x_i = i\}$, consider the model

$$(8.23) \quad \text{logit}[P(Y \leq j|X = x_i)] = \alpha_j + \beta x_i.$$

- a. Show that $\text{logit}[P(Y \leq j|X = x_{i+1})] - \text{logit}[P(Y \leq j|X = x_i)] = \beta$ is a log cumulative odds ratio for the 2×2 table consisting of rows i and $i + 1$ and the binary response having cutpoint following category j . Thus, (8.23) is a *uniform association model* in cumulative odds ratios.
- b. Show that (i) residual df = $IJ - I - J$ and (ii) $\beta = 0$ corresponds to independence of X and Y .
- c. Using the same linear predictor but with adjacent-categories logits, show that uniform association applies to the local odds ratios (2.10).

8.39 A cumulative link model for an $I \times J$ table with a qualitative predictor is

$$G^{-1}[P(Y \leq j)] = \alpha_j + \mu_i, \quad i = 1, \dots, I, \quad j = 1, \dots, J - 1.$$

Show that (a) residual df = $(I - 1)(J - 2)$, (b) independence corresponds to $\mu_1 = \dots = \mu_I$, (c) the test of independence has df = $I - 1$, and (d) the rows are stochastically ordered on Y .

8.40 Prove factorization (8.14) for the multinomial distribution.

8.41 A response scale has the categories (strongly agree, mildly agree, mildly disagree, strongly disagree, don't know). One model uses a logistic part for $P(\text{don't know})$ and a separate ordinal part for the ordered categories conditional on response in one of those categories. Explain how to construct a likelihood function to do this simultaneously.

8.42 For cumulative link model (8.7), show that for $1 \leq j < k \leq J - 1$, $P(Y \leq k|\mathbf{x}) = P(Y \leq j|\mathbf{x}^*)$, where \mathbf{x}^* is obtained by increasing the i th component of \mathbf{x} by $(\alpha_k - \alpha_j)/\beta_i$. Interpret.

8.43 When X and Y are ordinal, explain how to test conditional independence by allowing a different trend in each partial table. [Hint: Generalize model (8.17) by replacing β by β_k .]

8.44 Consider equation (8.21) and the condition of independence from irrelevant alternatives. Explain why this condition does not hold for the multinomial probit model.

8.45 For a Bayesian analysis, explain why the posterior $P(\beta \leq 0)$ is analogous to the frequentist P -value for $H_a: \beta > 0$.

8.46 After fitting a cumulative logit model of proportional odds form, what might you do to check the model (a) as part of a frequentist analysis and (b) as part of a Bayesian analysis?

¹Obtained using PROC LOGISTIC in SAS.

CHAPTER 9

Loglinear Models for Contingency Tables

In Section 4.3 we introduced loglinear models as generalized linear models (GLMs) using the log link function with a Poisson response. A common use is modeling for contingency tables. The models specify the joint distribution among the categorical variables that are cross-classified to form the table. They are used to analyze association and interaction patterns among those variables. The models specify how the expected cell counts depend on levels of the categorical variables for that cell as well as associations and interactions among those variables.

We present loglinear models in Section 8.1 for two-way contingency tables, in Sections 8.2 and 8.3 for three-way tables, and in Section 8.4 for multiway tables. Loglinear models are of use primarily when at least two variables are response variables. With a single categorical response, it is simpler and more natural to use logistic regression models. When one variable is treated as a response and the others as explanatory variables, logistic models for that response variable are equivalent to certain loglinear models. Section 8.5 presents this connection. In Sections 8.6 and 8.7 we discuss loglinear model fitting.

9.1 LOGLINEAR MODELS FOR TWO-WAY TABLES

Consider an $I \times J$ contingency table that cross-classifies a multinomial sample of n subjects on two categorical responses. For cell probabilities $\{\pi_{ij}\}$, the expected frequencies are $\{\mu_{ij} = n\pi_{ij}\}$. Loglinear model formulas use $\{\mu_{ij}\}$ rather than $\{\pi_{ij}\}$, so they also apply with Poisson sampling for $N = IJ$ independent cell counts $\{Y_{ij}\}$ having $\{\mu_{ij} = E(Y_{ij})\}$. In either case we denote the observed cell counts by (n_{ij}) .

9.1.1 Independence Model for a Two-Way Table

Under statistical independence, the $\{\mu_{ij}\}$ have the structure

$$\mu_{ij} = \mu\alpha_i \beta_j$$

(Section 4.3.7). For multinomial sampling, for instance, $\mu_{ij} = n\pi_{i+} \pi_{+j}$. Denote the row variable by X and the column variable by Y . The formula expressing independence is multiplicative. Thus, $\log \mu_{ij}$ has additive form¹

$$(9.1) \quad \log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$$

for a row effect λ_i^X and a column effect λ_j^Y . This is the *loglinear model of independence*. Identifiability requires constraints such as $\lambda_i^X = \lambda_j^Y = 0$ or $\sum_i \lambda_i^X = \sum_j \lambda_j^Y = 0$.

The ML fitted values are $\{\hat{\mu}_{ij} = n_{i+}n_{+j}/n\}$, the estimated expected frequencies for chi-squared tests of independence. The tests using X^2 and G^2 (Section 3.2.1) are goodness-of-fit tests of this loglinear model.

9.1.2 Interpretation of Loglinear Model Parameters

Loglinear models for contingency tables are GLMs that treat the N cell counts as independent observations of a Poisson random component. Loglinear GLMs identify the data as the N cell counts rather than the individual classifications of the n subjects. The expected cell counts link to the explanatory terms using the log link. As (9.1) illustrates, of the cross-classified variables, the model does not distinguish between response and explanatory variables. It treats both jointly as responses, modeling $\{\mu_{ij}\}$ for combinations of their levels. To interpret parameters, however, it is helpful to treat the variables asymmetrically.

We illustrate with the independence model for $I \times 2$ tables. In row i , the logit equals

$$\begin{aligned} \text{logit}[P(Y = 1|X = i)] &= \log \frac{P(Y = 1|X = i)}{P(Y = 2|X = i)} \\ &= \log \frac{\mu_{i1}}{\mu_{i2}} = \log \mu_{i1} - \log \mu_{i2} \\ &= (\lambda + \lambda_i^X + \lambda_1^Y) - (\lambda + \lambda_i^X + \lambda_2^Y) = \lambda_1^Y - \lambda_2^Y. \end{aligned}$$

The final term does not depend on i ; that is, $\text{logit}[P(Y = 1|X = i)]$ is identical at each level of X . Thus, independence implies a model of form $\text{logit}[P(Y = 1|X = i)] = \alpha$. In each row, the odds of response in column 1 equal $\exp(\alpha) = \exp(\lambda_1^Y - \lambda_2^Y)$. When $\lambda_1^Y = \lambda_2^Y$, the model simplifies to $\log \mu_{ij} = \lambda + \lambda_i^Y$ and there is *equiprobability*, with $P(Y = 1|X = i) = P(Y = 2|X = i)$ in each row.

An analogous property holds when $J > 2$. Differences between two parameters for a variable relate to the log odds of making one response, relative to the other, on that variable.

9.1.3 Saturated Model for a Two-Way Table

Statistically dependent variables satisfy a more complex loglinear model,

$$(9.2) \quad \log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}.$$

The $\{\lambda_{ij}^{XY}\}$ are association terms that reflect deviations from independence. The right-hand side of (9.2) resembles the formula for cell means in two-way ANOVA, allowing interaction. The $\{\lambda_{ij}^{XY}\}$ represent interactions between X and Y , whereby the effect of one variable on μ_{ij} depends on the level of the other. The independence model (9.1) results when all $\lambda_{ij}^{XY} = 0$.

With constraints $\lambda_I^X = \lambda_J^Y = 0$ in (9.1) and (9.2), $\{\lambda_i^X\}$ and $\{\lambda_j^Y\}$ are, equivalently, coefficients of indicator variables for the first $(I - 1)$ categories of X and for the first $(J - 1)$ categories of Y . Thus, λ_{ij}^{XY} is the coefficient of the product of indicator variables for λ_i^X and λ_j^Y . Since there are $(I - 1)(J - 1)$ such cross-products, $\lambda_{Ij}^{XY} = \lambda_{iJ}^{XY} = 0$, and only $(I - 1)(J - 1)$ of these parameters are nonredundant. Tests of independence analyze whether these $(I - 1)(J - 1)$ parameters equal zero, so they have residual df = $(I - 1)(J - 1)$.

The number of parameters in model (9.2) equals $1 + (I - 1) + (J - 1) + (I - 1)(J - 1) = IJ$, the number of cells. Hence, this model describes perfectly any $\{\mu_{ij} > 0\}$ (see Exercise 9.16). The ML fitted values are $\{\hat{\mu}_{ij} = n_{ij}\}$. It is the most general model for two-way contingency tables, the *saturated model*. For it, direct relationships exist between log odds ratios and $\{\lambda_{ij}^{XY}\}$. For instance, for 2×2 tables,

$$\begin{aligned} \log \theta &= \log \frac{\mu_{11} \mu_{22}}{\mu_{12} \mu_{21}} = \log \mu_{11} + \log \mu_{22} - \log \mu_{12} - \log \mu_{21} \\ &= (\lambda + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}) + (\lambda + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}) \\ &\quad - (\lambda + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}) - (\lambda + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY}) \\ (9.3) \quad &= \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}. \end{aligned}$$

Thus, $\{\lambda_{ij}^{XY}\}$ determine the association.

In practice, unsaturated models are preferable, since their fit smooths the sample data and has simpler interpretations. For tables with at least three variables, unsaturated models can include association terms. Then, loglinear models are more commonly used to describe associations (through two-factor terms) than to describe odds (through single-factor terms).

9.1.4 Alternative Parameter Constraints

As with the independence model, the parameter constraints for the saturated model are arbitrary. Instead of setting all $\lambda^{XY}_{Ij} = \lambda^{XY}_{iJ} = 0$, we could set $\sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = 0$ for all i and j . Different software uses different constraints. What is unique and estimable are contrasts such as $\lambda^{XY}_{11} + \lambda^{XY}_{22} - \lambda^{XY}_{12} - \lambda^{XY}_{21}$ in (9.3) that determine odds ratios.

For instance, suppose that a log odds ratio equals 2.0 in a 2×2 table. With the first set of constraints, 2.0 is the coefficient of a product of an indicator variable indicating the first category of X and an indicator variable indicating the first category of Y . With it, $\lambda^{XY}_{11} = 2.0$ and $\lambda^{XY}_{12} = \lambda^{XY}_{21} = \lambda^{XY}_{22} = 0$. For sum-to-zero constraints, $\lambda^{XY}_{11} = \lambda^{XY}_{22} = 0.5$, $\lambda^{XY}_{12} = \lambda^{XY}_{21} = -0.5$. For either set, the log odds ratio (9.3) equals 2.0.

9.1.5 Hierarchical Versus Nonhierarchical Models

Like other models in this book, model (9.2) is *hierarchical*. This means that the model includes all lower-order terms composed from variables contained in a higher-order term. When the model contains λ^{XY}_{ij} , it also contains λ^X_i and λ^Y_j . A reason for including lower-order terms is that, otherwise, the statistical significance and the interpretation of a higher-order term depends on how variables are coded. This is undesirable, and with hierarchical models the same results occur no matter how variables are coded.

An example of a nonhierarchical model is

$$\log \mu_{ij} = \lambda + \lambda^X_i + \lambda^{XY}_{ij}.$$

This model permits association but forces unnatural behavior of expected frequencies, with the pattern depending on constraints used for parameters. For instance, with constraints whereby parameters are zero at the last level, $\log \mu_{Ij} = \lambda$ in every column. Nonhierarchical models are rarely sensible in practice. Using them is analogous to using ANOVA or regression models with interaction terms but without the corresponding main effects.

When a model has two-factor terms, interpretations focus on them rather than on the single-factor terms. By analogy with two-way ANOVA with two-factor interaction, it can be misleading to report main effects. The estimates of the main-effect terms depend on the coding scheme used for the higher-order effects, and the interpretation also depends on that scheme (see Exercise 9.16). Normally, we restrict our attention to the highest-order terms for a variable, as we illustrate in Section 9.2.

9.1.6 Multinomial Models for Cell Probabilities

Conditional on the sum n of the cell counts, Poisson loglinear models for $\{\mu_{ij}\}$ become multinomial models for cell probabilities $\{\pi_{ij} = \mu_{ij}/(\sum_a \sum_b \mu_{ab})\}$. To illustrate, for the saturated model,

$$(9.4) \quad \pi_{ij} = \frac{\exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY})}{\sum_a \sum_b \exp(\lambda + \lambda_a^X + \lambda_b^Y + \lambda_{ab}^{XY})}.$$

The λ intercept parameter cancels in the multinomial model (9.4). This parameter relates to the total sample size, which is random in the Poisson model but not in the multinomial model. So, the saturated multinomial model has $I J - 1$ parameters, representing the usual constraint for probabilities, $\sum_i \sum_j \pi_{ij} = 1$.

9.2 LOGLINEAR MODELS FOR INDEPENDENCE AND INTERACTION IN THREE-WAY TABLES

In Section 2.3 we introduced structure for three-way contingency tables, such as conditional independence and homogeneous association. Loglinear models for three-way tables describe these independence and association patterns.

9.2.1 Types of Independence

A three-way $I \times J \times K$ cross-classification of response variables X , Y , and Z has several potential types of independence. The models apply to a multinomial distribution with cell probabilities $\{\pi_{ijk}\}$ having $\sum_i \sum_j \sum_k \pi_{ijk} = 1.0$ and also to Poisson sampling with means $\{\mu_{ijk}\}$.

Mutual independence: The three variables are *mutually independent* when

$$(9.5) \quad \pi_{ijk} = \pi_{i++} \pi_{+j+} \pi_{++k} \quad \text{for all } i, j, \text{ and } k.$$

For expected frequencies $\{\mu_{ijk}\}$, mutual independence has loglinear form

$$(9.6) \quad \log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z.$$

Joint independence: Variable Y is *jointly independent* of X and Z when

$$(9.7) \quad \pi_{ijk} = \pi_{i+k} \pi_{+j+} \quad \text{for all } i, j, \text{ and } k.$$

This is ordinary two-way independence between Y and a variable composed of the IK combinations of levels of X and Z . The loglinear model is

$$(9.8) \quad \log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}.$$

Similarly, X could be jointly independent of Y and Z , or Z could be jointly independent of X and Y . Mutual independence (9.5) implies joint independence of any one variable from the other two.

Conditional independence: Categorical variables X and Y are *conditionally independent*, given Z , when independence holds for each partial table within which Z is fixed. That is, if $\mu_{ij|k} = P(X=i, Y=j|Z=k)$, then

$$\pi_{ij|k} = \pi_{i+k} \pi_{+j+k} \quad \text{for all } i, j, \text{ and } k.$$

For joint probabilities over the entire table, equivalently

$$(9.9) \quad \pi_{ijk} = \pi_{i+k} \pi_{+jk} / \pi_{++k} \quad \text{for all } i, j, \text{ and } k.$$

Conditional independence of X and Y , given Z , has loglinear model form

$$(9.10) \quad \log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.$$

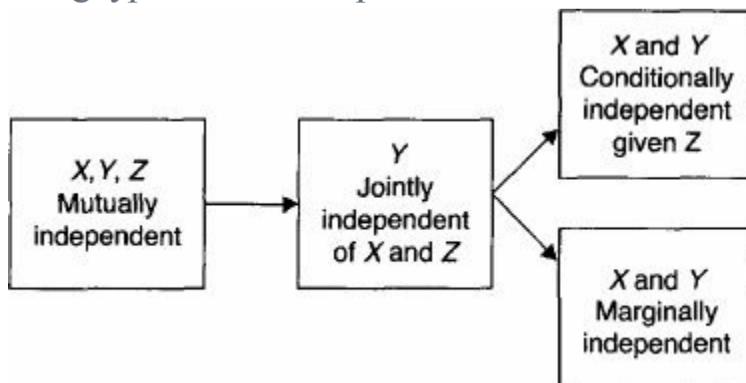
This is a weaker condition than mutual or joint independence. Mutual independence implies that Y is jointly independent of X and Z , which itself implies that X and Y are conditionally independent. [Table 9.1](#) summarizes these three types of independence.

Table 9.1 Loglinear Independence Models for Three-Dimensional Tables

Model Formula	Probabilistic Form for π_{ijk}	Association Terms in Loglinear Model	Interpretation
(9.6)	$\pi_{i++} \pi_{+j+} \pi_{++k}$	None	Mutual independence of X , Y , Z
(9.8)	$\pi_{i+k} \pi_{+j+}$	λ_{ik}^{XZ}	Joint independence of Y from X and Z
(9.10)	$\pi_{i+k} \pi_{+jk} / \pi_{++k}$	$\lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$	Conditional independence of X and Y , given Z

Recall that conditional associations can be quite different from marginal associations (Section 2.3.2). For instance, conditional independence does not imply marginal independence. Conditional independence and marginal independence both hold when one of the stronger types of independence studied above applies. [Figure 9.1](#) summarizes relationships among the four types of independence.

Figure 9.1 Relationships among types of XY independence.



9.2.2 Homogeneous Association and Three-Factor Interaction

Loglinear models (9.6), (9.8), and (9.10) have three, two, and one pair of conditionally independent variables, respectively. In the latter two models, the doubly subscripted terms (such as λ_{ij}^{XY}) pertain to conditionally dependent variables. A model that permits all three pairs to be conditionally dependent is

$$(9.11) \quad \log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.$$

From exponentiating both sides, the cell probabilities have form

$$\pi_{ijk} = \psi_{ij}\phi_{jk}\omega_{ik}.$$

No closed-form expression exists for the three components in terms of margins of $\{\pi_{ijk}\}$ except in certain special cases (see Note 10.2).

For this model, in the next section we show that conditional odds ratios between any two variables are identical at each category of the third variable. That is, each pair has *homogeneous association*, as first defined for $2 \times 2 \times K$ tables in Section 2.3.5. Model (9.11) is called the loglinear model of *homogeneous association* or of *no three-factor interaction*.

The general loglinear model for a three-way table is

$$(9.12) \quad \log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}.$$

With indicator variables, λ_{ijk}^{XYZ} is the coefficient of the product of the i th indicator variable for X , j th indicator variable for Y , and k th indicator variable for Z . The total number of nonredundant parameters is which is the total number of cell counts. This model has as many parameters as observations and is saturated. It describes all possible $\{\mu_{ijk} > 0\}$. Each pair of variables may be conditionally dependent, and an odds ratio for any pair may vary across categories of the third variable.

$$1 + (I - 1) + (J - 1) + (K - 1) + (I - 1)(J - 1) + (I - 1)(K - 1) \\ + (J - 1)(K - 1) + (I - 1)(J - 1)(K - 1) = IJK,$$

Setting certain parameters equal to zero in (9.12) yields the models introduced previously. Table 9.2 lists some of these models. To ease referring to models, this table assigns to each model a symbol that lists the highest-order term(s) for each variable. For instance, the model (9.10) of conditional independence between X and Y has symbol (XZ, YZ) , since its highest-order terms are λ_{ik}^{XZ} and λ_{jk}^{YZ} . In the notation we used for logistic models in Sections 6.1 and 8.1.2 this stands for $(X^*Z + Y^*Z)$, which is itself shorthand for notation $[X + Y + Z + (X \cdot Z) + (Y \cdot Z)]$ that has the main effects as well as two interactions.

Table 9.2 Loglinear Models for Three-Dimensional Tables

Loglinear Model Formula	Symbol
$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$	(X, Y, Z)
$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$	(XY, Z)
$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$	(XY, YZ)
$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ}$	(XY, YZ, XZ)
$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ} + \lambda_{ijk}^{XYZ}$	(XYZ)

9.2.3 Interpretation of Loglinear Model Parameters

Interpretations of loglinear model parameters use their highest-order terms. For instance, interpretations for model (9.11) use the two-factor terms to describe conditional odds ratios. At a fixed level k of Z , the *conditional association* between X and Y uses $(I - 1)(J - 1)$ odds ratios, such as the local odds ratios

$$(9.13) \quad \theta_{ij(k)} = \frac{\pi_{ijk} \pi_{i+1,j+1,k}}{\pi_{i,j+1,k} \pi_{i+1,j,k}}, \quad 1 \leq i \leq I - 1, \quad 1 \leq j \leq J - 1.$$

Similarly, $(I - 1)(K - 1)$ odds ratios $\{\theta_{ij(k)}\}$ describe XZ conditional association, and $(J - 1)(K - 1)$ odds ratios $\{\theta_{(i)jk}\}$ describe YZ conditional association. Loglinear models have characterizations using constraints on conditional odds ratios. For instance, conditional independence of X and Y is equivalent to $\{\theta_{ik(k)} = 1, i = 1, \dots, I - 1, j = 1, \dots, J - 1, k = 1, \dots, K\}$.

The two-factor parameters relate directly to the conditional odds ratios. To illustrate, substituting (9.11) for model (XY, XZ, YZ) into $\log \theta_{ij(k)}$ yields

$$(9.14) \quad \log \theta_{ij(k)} = \log \frac{\mu_{ijk} \mu_{i+1,j+1,k}}{\mu_{i+1,jk} \mu_{i,j+1,k}} = \lambda_{ij}^{XY} + \lambda_{i+1,j+1}^{XY} - \lambda_{i,j+1}^{XY} - \lambda_{i+1,j}^{XY}.$$

Since the right-hand side is the same for all k , an absence of three-factor interaction is equivalent to

$$\theta_{ij(1)} = \theta_{ij(2)} = \dots = \theta_{ij(k)} \quad \text{for all } i \text{ and } j.$$

The same argument for the other conditional odds ratios shows that model (XY, XZ, YZ) is also equivalent to

$$\theta_{i(1)k} = \theta_{i(2)k} = \dots = \theta_{i(J)k} \quad \text{for all } i \text{ and } k,$$

and to

$$\theta_{(1)jk} = \theta_{(2)jk} = \dots = \theta_{(I)jk} \quad \text{for all } j \text{ and } k.$$

Any model not having the three-factor interaction term has a homogeneous association for each pair of variables.

When X and Y have two categories, only one nonredundant λ_{ij}^{XY} parameter occurs. Thus, expression (9.14) simplifies according to the constraints. By the same argument as in Section 9.1.3 for 2×2 tables, the conditional log odds ratio simplifies to λ_{11}^{XY} with indicator-variable constraints setting parameters at the second level of X or Y equal to 0.

The λ_{ijk}^{XYZ} term in the general model (9.12) refers to three-factor interaction. It describes how the odds ratio between two variables changes across categories of the third. We illustrate for $2 \times 2 \times 2$ tables. By direct substitution of the general model formula,

$$\begin{aligned} \log \frac{\theta_{11(1)}}{\theta_{11(2)}} &= \log \frac{(\mu_{111} \mu_{221}) / (\mu_{121} \mu_{211})}{(\mu_{112} \mu_{222}) / (\mu_{122} \mu_{212})} \\ &= (\lambda_{111}^{XYZ} + \lambda_{221}^{XYZ} - \lambda_{121}^{XYZ} - \lambda_{211}^{XYZ}) \\ &\quad - (\lambda_{112}^{XYZ} + \lambda_{222}^{XYZ} - \lambda_{122}^{XYZ} - \lambda_{212}^{XYZ}). \end{aligned}$$

Only one parameter is nonredundant. For constraints setting the second-category parameters equal to 0, this log ratio of odds ratios equals λ_{111}^{XYZ} . When $\lambda_{111}^{XYZ} = 0$, $\theta_{11(1)} = \theta_{11(2)}$, giving homogeneous XY association.

9.2.4 Example: Alcohol, Cigarette, and Marijuana Use

[Table 9.3](#) refers to a survey by the Wright State University School of Medicine and the United Health Services in Dayton, Ohio. The survey asked 2276 students in their final year of high school in a nonurban area near Dayton, Ohio, whether they had ever used alcohol, cigarettes, or marijuana. Denote the variables in this $2 \times 2 \times 2$ table by A for alcohol use, C for cigarette use, and M for marijuana use.

Table 9.3 Alcohol, Cigarette, and Marijuana Use for High School Seniors

Alcohol Use	Cigarette Use	Marijuana Use	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

Source: Data courtesy of Harry Khamis, Wright State University.

Section 9.7 covers the fitting of loglinear models. For now, we emphasize interpretation. [Table 9.4](#) shows fitted values for several loglinear models. The fit for model (AC, AM, CM) is close to the observed data, which are the fitted values for the saturated model (ACM) . The other models fit poorly.

Table 9.4 Fitted Values for Loglinear Models Applied to [Table 9.3](#)

Alcohol Use (A)	Cigarette Use (C)	Marijuana Use (M)	Loglinear Model				
			(A,C,M)	(AC,M)	(AM,CM)	(AC,AM,CM)	(ACM)
Yes	Yes	Yes	540.0	611.2	909.24	910.4	911
		No	740.2	837.8	438.84	538.6	538
	No	Yes	282.1	210.9	45.76	44.6	44
		No	386.7	289.1	555.16	455.4	456
No	Yes	Yes	90.6	19.4	4.76	3.6	3
		No	124.2	26.6	142.16	42.4	43
	No	Yes	47.3	118.5	0.24	1.4	2
		No	64.9	162.5	179.84	279.6	279

[Table 9.5](#) illustrates model association patterns by presenting estimated conditional and marginal odds ratios. For example, the entry 1.0 for the AC conditional association for the model (AM, CM) of AC conditional independence is the common value of the AC fitted odds ratios at the two levels of M ,

$$1.0 = \frac{909.24 \times 0.24}{45.76 \times 4.76} = \frac{438.84 \times 179.84}{555.16 \times 142.16}.$$

The entry 2.7 for the AC marginal association for this model is the odds ratio for the marginal AC fitted table. The odds ratios for the observed data are those reported for the saturated model (ACM) .

[Table 9.5](#) shows that estimated conditional odds ratios equal 1.0 for each pairwise term not appearing in a model, such as the AC association in model (AM, CM) . For that model, the estimated marginal AC odds ratio differs from 1.0, since conditional independence does not imply marginal independence. Some models have conditional associations that are necessarily the same as the corresponding marginal associations. In Section 10.1.3 we present a condition guaranteeing this.

Table 9.5 Estimated Odds Ratios for Loglinear Models in [Table 9.4](#)

Model	Conditional Association			Marginal Association		
	<i>AC</i>	<i>AM</i>	<i>CM</i>	<i>AC</i>	<i>AM</i>	<i>CM</i>
(<i>A, C, M</i>)	1.0	1.0	1.0	1.0	1.0	1.0
(<i>AC, M</i>)	17.7	1.0	1.0	17.7	1.0	1.0
(<i>AM, CM</i>)	1.0	61.9	25.1	2.7	61.9	25.1
(<i>AC, AM, CM</i>)	7.8	19.8	17.3	17.7	61.9	25.1
(<i>ACM</i>) level 1	13.8	24.3	17.5	17.7	61.9	25.1
(<i>ACM</i>) level 2	7.7	13.5	9.7			

Model (*AC, AM, CM*) permits all pairwise associations but maintains homogeneous odds ratios between two variables at each level of the third. The *AC* fitted conditional odds ratios for this model equal 7.8. We can calculate this odds ratio using the model's fitted values at either level of *M*, or [from (9.14)] using $\exp(\hat{\lambda}_{11}^{AC} + \hat{\lambda}_{22}^{AC} - \hat{\lambda}_{12}^{AC} - \hat{\lambda}_{21}^{AC})$.

[Table 9.5](#) shows that estimated odds ratios are highly dependent on the model. So, good model selection is crucial. An estimate from this table is informative only to the extent that its model fits well. In the next section we discuss goodness of fit.

9.3 INFERENCE FOR LOGLINEAR MODELS

A good-fitting loglinear model provides a basis for describing and making inferences about associations among categorical responses. Standard methods apply for checking fit and making inference about model parameters.

9.3.1 Chi-Squared Goodness-of-Fit Tests

As usual, X^2 and G^2 test whether a model holds by comparing cell fitted values to observed counts. For loglinear models, df equals the number of cell counts minus the number of model parameters.

For the student survey data ([Table 9.3](#)), [Table 9.6](#) shows results of testing fit for several models. Models that lack any association term fit poorly. The model (AC , AM , CM) that has all pairwise associations fits well ($P = 0.54$). It is suggested by other criteria also, such as minimizing AIC (Section 6.1.6).

Table 9.6 Goodness-of-Fit Tests for Loglinear Models in [Table 9.4](#)

Loglinear Model	G^2	X^2	df	P -value ^a	AIC
(A, C, M)	1286.0	1411.4	4	< 0.001	1343.1
(A, CM)	534.2	505.6	3	< 0.001	593.3
(C, AM)	939.6	824.2	3	< 0.001	998.6
(M, AC)	843.8	704.9	3	< 0.001	902.9
(AC, AM)	497.4	443.8	2	< 0.001	558.4
(AC, CM)	92.0	80.8	2	< 0.001	153.1
(AM, CM)	187.8	177.6	2	< 0.001	248.8
(AC, AM, CM)	0.4	0.4	1	0.54	63.4
(ACM)	0.0	0.0	0	—	65.0

^a P -value for G^2 deviance statistic.

9.3.2 Inference about Conditional Associations

Tests about conditional associations compare loglinear models. The likelihood-ratio statistic $-2(L_0 - L_1)$ is identical to the difference $G^2(M_0|M_1) = G^2(M_0) - G^2(M_1)$ between deviances for models without that term and with it. For model (XY, XZ, YZ) , consider the hypothesis of XY conditional independence. This is $H_0: \lambda_{ij}^{XY} = 0$ for the $(I-1)(J-1)$ association parameters relating X and Y . The test statistic is $G^2(XZ, YZ) - G^2(XY, XZ, YZ)$, with $df = (I-1)(J-1)$. This has the same purpose as the generalized CMH and model-based tests for nominal variables presented in Section 8.4.

For instance, the test of conditional independence between alcohol use and cigarette smoking compares model (AM, CM) with the alternative (AC, AM, CM) . The test statistic is

$$G^2[(AM, CM)|(AC, AM, CM)] = 187.8 - 0.4 = 187.4,$$

with $df = 2 - 1 = 1$ ($P < 0.001$). The statistics comparing (AC, CM) and (AC, AM) with (AC, AM, CM) also provide strong evidence of AM and CM conditional associations. In further analyses of the data, we use model (AC, AM, CM) .

With large sample sizes, statistically significant effects can be weak and practically unimportant. A more relevant concern is whether the associations are strong enough to be of interest. Confidence intervals are more useful than tests for assessing this. [Table 9.7](#) shows output from fitting model (AC, AM, CM) with parameters in the last row and in the last column equal to zero, such as by using $(1, 0)$ indicator variables for each classification. Consider the conditional AC odds ratio, assuming model (AC, AM, CM) . [Table 9.7](#) reports $\hat{\lambda}_{11}^{AC} = 2.054$, with $SE = 0.174$. For these constraints, this is the estimated conditional log odds ratio. A 95% Wald confidence interval for the true conditional AC odds ratio is $\exp[2.054 \pm 1.96(0.174)]$, or $(5.5, 11.0)$. Strong positive association exists between cigarette use and alcohol use, for both users and nonusers of marijuana.

Table 9.7 Software Output (Based on SAS) for Homogeneous Association Model Fitted to [Table 9.3](#)

Criteria For Assessing Goodness Of Fit				
Criterion	DF	Value	Value/DF	
Deviance	1	0.3740	0.3740	
Pearson Chi-Square	1	0.4011	0.4011	
Standard Wald				
Parameter	Estimate	Error	Chi-Square	Pr > ChiSq
Intercept	5.6334	0.0597	8903.96	<.0001
a	1	0.4877	41.44	<.0001
c	1	-1.8867	134.47	<.0001
m	1	-5.3090	124.82	<.0001
a*m	1 1	2.9860	41.29	<.0001
a*c	1 1	2.0545	139.32	<.0001
c*m	1 1	2.8479	302.14	<.0001
LR Statistics				
Source	DF	Chi-Square	Pr > ChiSq	
a*m	1	91.64	<.0001	
a*c	1	187.38	<.0001	
c*m	1	497.00	<.0001	

For model (AC, AM, CM) , the 95% Wald confidence intervals are $(8.0, 49.2)$ for the AM conditional odds ratio and $(12.5, 23.8)$ for the CM conditional odds ratio. The intervals are wide, but these associations also are strong. [Table 9.5](#) showed that estimated marginal associations are even stronger. Controlling for outcome on one response moderates the association somewhat between the other two.

The analyses in this section pertain to associations. A different analysis pertains to comparing single-variable marginal distributions, for instance, to determine if students used cigarettes more than alcohol or marijuana. That type of analysis is presented in Section 11.1.

9.4 LOGLINEAR MODELS FOR HIGHER DIMENSIONS

Loglinear models for three-way tables extend readily to multiway tables. As the number of dimensions increases, some complications arise. One is the increase in the number of possible association and interaction terms, making model selection more difficult. Another is the increase in number of cells. In Section 10.6 we show that this can cause difficulties with existence of estimates and appropriateness of some large-sample theory.

9.4.1 Models for Four-Way Contingency Tables

We illustrate models for higher dimensions using a four-way table with variables W, X, Y , and Z . Interpretations are simplest when the model has no three-factor interaction terms, so that each pairwise association is homogeneous. Such models are special cases of

$$\begin{aligned}\log \mu_{hijk} = & \lambda + \lambda_h^W + \lambda_i^X + \lambda_j^Y + \lambda_k^Z \\ & + \lambda_{hi}^{WX} + \lambda_{hj}^{WY} + \lambda_{hk}^{WZ} + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ},\end{aligned}$$

denoted by (WX, WY, WZ, XY, XZ, YZ) . Each pair of variables is conditionally dependent, with the same odds ratios at each combination of categories of the other two variables. An absence of a two-factor term implies conditional independence, given the other two variables.

A variety of models exhibit three-factor interaction. A model could contain any of WXY, WXZ, WYZ , or XYZ terms. For model (WXY, WZ, XZ, YZ) , each pair of variables is conditionally dependent, but at each level of Z the WX association, the WY association, and the XY association may vary across categories of the remaining variable. The conditional association between Z and another variable is homogeneous. The saturated model contains all the three-factor terms plus a four-factor interaction term.

9.4.2 Example: Automobile Accidents and Seat-Belt Use

[Table 9.8](#) summarizes observations of 68,694 passengers in autos and light trucks involved in accidents one year in the state of Maine. The table classifies passengers by gender (G), location of accident (L), seat-belt use (S), and injury (I). [Table 9.8](#) reports the sample proportion of passengers who were injured. For each GL combination, the proportion of injuries was about halved for passengers wearing seat belts.

Table 9.8 Loglinear Models for Injury, Seat-Belt Use, Gender, and Location^a

Gender	Location	Seat Belt Use	Injury Observed		Injury (GI, GL, GS, IL, IS, LS)		Sample Proportion
			No	Yes	No	Yes	
Female	Urban	No	7,287	996	7,166.4	993.0	7,273.2
		Yes	11,587	759	11,748.3	721.3	11,632.6
	Rural	No	3,246	973	3,353.8	988.8	3,254.7
		Yes	6,134	757	5,985.5	781.9	6,093.5
Male	Urban	No	10,381	812	10,471.5	845.1	10,358.9
		Yes	10,969	380	10,837.8	387.6	10,959.2
	Rural	No	6,123	1,084	6,045.3	1,038.1	6,150.2
		Yes	6,693	513	6,811.4	518.2	6,697.6

^a G , gender; I , injury; L , location; S , seat-belt use.

Source: Data courtesy of Cristanna Cook, Medical Care Development, Augusta, Maine.

[Table 9.9](#) displays tests of fit for several loglinear models. To investigate the complexity of model needed, we consider models (G, I, L, S) , (GI, GL, GS, IL, IS, LS) , and (GIL, GIS, GLS, ILS) having all terms of varying complexity. Model (G, I, L, S) of mutual independence fits very poorly. Model (GI, GL, GS, IL, IS, LS) fits much better but still has a lack of fit ($P < 0.001$). Model (GIL, GIS, GLS, ILS) fits well ($G^2 = 1.33$, $df = 1$) but is complex and difficult to interpret. This suggests studying models more complex than (GI, GL, GS, IL, IS, LS) but simpler than (GIL, GIS, GLS, ILS) .

Table 9.9 Goodness-of-Fit Tests for Loglinear Models Fitted to [Table 9.8](#)

Model	G^2	df	P-Value	AIC
(G, I, L, S)	2792.77	11	<0.0001	2956.2
(GI, GL, GS, IL, IS, LS)	23.35	5	<0.001	198.8
(GIL, GIS, GLS, ILS)	1.33	1	0.25	184.8
(GIL, GS, IS, LS)	18.57	4	0.001	196.0
(GIS, GL, IL, LS)	22.85	4	<0.001	200.3
(GLS, GI, IL, IS)	7.46	4	0.11	184.9
(ILS, GI, GL, GS)	20.63	4	<0.001	198.1

First, however, we analyze model (GI, GL, GS, IL, IS, LS) , which focuses on pairwise associations. [Table 9.8](#) displays its fitted values. [Table 9.10](#) reports the model-based estimated conditional odds ratios. We can obtain them directly from parameter estimates; for instance, $0.44 = \exp(\hat{\lambda}_{11}^{IS} + \hat{\lambda}_{22}^{IS} - \hat{\lambda}_{12}^{IS} - \hat{\lambda}_{21}^{IS})$.

Table 9.10 Estimated Conditional Odds Ratios for Models for [Table 9.8](#)

Loglinear Model		
Odds Ratio	(GI, GL, GS, IL, IS, LS)	(GLS, GI, IL, IS)
GI	0.58	0.58
IL	2.13	2.13
IS	0.44	0.44
$GL S = (\text{no, yes})$	(1.23, 1.23)	(1.33, 1.17)
$GS L = (\text{urban, rural})$	(0.63, 0.63)	(0.66, 0.58)
$LS G = (\text{female, male})$	(1.09, 1.09)	(1.17, 1.03)

Since the sample size is large, the estimates of odds ratios are quite precise. For instance, the standard error of the estimated IS conditional log odds ratio of -0.814 is 0.028 . A 95% Wald confidence interval for the true odds ratio is $\exp[-0.814 \pm 1.96(0.028)]$ or $(0.42, 0.47)$. This model estimates that the odds of injury for passengers wearing seat belts were less than half the odds for passengers not wearing them, at each gender-by-location combination. The fitted odds ratios in [Table 9.10](#) also suggest that other factors being fixed, injury was more likely in rural than urban accidents and more likely for females than for males. The estimated odds that males used seat belts were 0.63 times the estimated odds for females, conditional on each combination of I and L categories.

Interpretations are more complex for models containing three-factor interaction terms. [Table 9.9](#) shows results of adding a single three-factor term to model (GI, GL, GS, IL, IS, LS) . Of the four possible models, (GLS, GI, IL, IS) fits best and has AIC essentially the same as the model (GIL, GIS, GLS, ILS) . [Table 9.8](#) also displays its fit. Considering that the sample size is very large, its G^2 value suggests that it fits quite well. For this model, each pair of variables is conditionally dependent, and at each category of I the association between any two of the others varies across categories of the remaining variable. For this model, it is inappropriate to interpret the GL , GS , and LS two-factor terms on their own. Since I does not occur in a three-factor interaction, the conditional odds ratio between I and each variable (see the top-right portion of [Table 9.10](#)) is the same at each combination of categories of the other two variables.

When a model has a three-factor interaction term but no term of higher order than that, we can study the interaction by calculating fitted odds ratios between two variables at each level of the third. We can do this at any levels of remaining variables not involved in the interaction. The bottom-right portion of [Table 9.10](#) illustrates this for model (GLS, GI, IL, IS) . For instance, the fitted GS odds ratio of 0.66 for (L = urban) refers to four fitted values for urban accidents, both the four with (injury = no) and the four with (injury = yes); for example, $0.66 = (7273.2 \times 10,959.2)/(11,632.6 \times 10,358.9)$.

9.4.3 Large Samples and Statistical Versus Practical Significance

Model (GLS, GI, IL, IS) seems to fit much better than (GI, GL, GS, IL, IS, LS) . The difference in G^2 values of $23.4 - 7.5 = 15.9$ has $df = 5 - 4 = 1$ ($P = 0.0001$). [Table 9.10](#) indicates, however, that the degree of three-factor interaction is weak. The fitted odds ratio between any two of G , L , and S is similar at both levels of the third variable. The significantly better fit of model (GLS, GI, IL, IS) reflects mainly the enormous sample size.

As in any test, a statistically significant effect need not be practically important. With huge samples, it is better to focus on estimation rather than hypothesis testing. For instance, a comparison of fitted odds ratios for the two models in [Table 9.10](#) suggests that the simpler model (GI, GL, GS, IL, IS, LS) is adequate for most purposes.

9.4.4 Dissimilarity Index

For a table of arbitrary dimension with cell counts $\{n_i = np_i\}$ and fitted values $\{\hat{\mu}_i = n_{\hat{\pi}i}\}$, the *dissimilarity index* (Gini 1914b)

$$\hat{\Delta} = \sum_i |n_i - \hat{\mu}_i|/2n = \sum_i |p_i - \hat{\pi}_i|/2$$

summarizes how far the model fit falls from the data. This index falls between 0 and 1, with larger values representing a poorer fit. It represents the proportion of sample cases that must move to different cells for the model to fit perfectly.

The dissimilarity index $\hat{\Delta}$ estimates a corresponding population index Δ describing model lack of fit. The value $\Delta = 0$ occurs when the model holds perfectly. In practice, this is unrealistic for unsaturated models, and $\Delta > 0$. The estimator $\hat{\Delta}$ helps study whether the lack of fit is important in a practical sense. When $\hat{\Delta} < 0.02$ or 0.03 , the sample data follow the model pattern quite closely, even though the model is not perfect.

For [Table 9.8](#), model (GI, GL, GS, IL, IS, LS) has $\hat{\Delta} = 0.008$, and model (GLS, GI, IL, IS) has $\hat{\Delta} = 0.003$. For either model, moving less than 1% of the data yields a perfect fit. The relatively large G^2 value for (GI, GL, GS, IL, IS, LS) indicated that it does not truly hold. Nevertheless, the small $\hat{\Delta}$ value suggests that, in practical terms, it fits decently.

When Δ is near 0, $\hat{\Delta}$ tends to overestimate Δ , substantially so for small n . Kuha and Firth (2011) provided an approximate variance for $\hat{\Delta}$ and studied ways to reduce its estimation bias.

9.5 LOGLINEAR—LOGISTIC MODEL CONNECTION

Loglinear models treat categorical response variables symmetrically, focusing on associations and interactions in their joint distribution. Logistic models, by contrast, describe how a single categorical response depends on explanatory variables. The model types seem distinct, but connections exist between them. For a loglinear model, forming logits on one response helps to interpret the model. Moreover, logistic models with categorical explanatory variables have equivalent loglinear models (Bishop 1969).

9.5.1 Using Logistic Models to Interpret Loglinear Models

To understand implications of a loglinear model formula, it can help to form a logit on one variable. We illustrate with the loglinear model (XY, XZ, YZ) . When Y is binary, its logit is

$$\begin{aligned}\log \frac{P(Y = 1|X = i, Z = k)}{P(Y = 2|X = i, Z = k)} &= \log \frac{\mu_{i1k}}{\mu_{i2k}} = \log \mu_{i1k} - \log \mu_{i2k} \\ &= (\lambda + \lambda_i^X + \lambda_1^Y + \lambda_k^Z + \lambda_{i1}^{XY} + \lambda_{ik}^{XZ} + \lambda_{1k}^{YZ}) \\ &\quad - (\lambda + \lambda_i^X + \lambda_2^Y + \lambda_k^Z + \lambda_{i2}^{XY} + \lambda_{ik}^{XZ} + \lambda_{2k}^{YZ}) \\ &= (\lambda_1^Y - \lambda_2^Y) + (\lambda_{i1}^{XY} - \lambda_{i2}^{XY}) + (\lambda_{1k}^{YZ} - \lambda_{2k}^{YZ}).\end{aligned}$$

The first parenthetical term is a constant, not depending on i or k . The second parenthetical term depends on the category i of X . The third parenthetical term depends on the category k of Z . This logit has the additive form

$$(9.15) \text{ logit}[P(Y = 1|X = i, Z = k)] = \alpha + \beta_i^X + \beta_k^Z.$$

Using the notation summarizing logistic models by their predictors, we denote it by $(X + Z)$.

In Section 5.4.1 we discussed this logistic model. When Y is binary, the loglinear model (XY, XZ, YZ) is equivalent to it. The λ_{ik}^{XZ} terms for association among explanatory variables cancel in the difference in logarithms that the logit defines. The logistic model does not study this association.

9.5.2 Example: Auto Accidents and Seat-Belts Revisited

For the Maine auto accidents ([Table 9.8](#)), in Section 9.4.2 we showed that the loglinear model (*GLS*, *GI*, *LI*, *IS*),

$$\begin{aligned}\log \mu_{g\ell s} = & \lambda + \lambda_g^G + \lambda_i^I + \lambda_\ell^L + \lambda_s^S + \lambda_{gi}^{GI} + \lambda_{g\ell}^{GL} + \lambda_{gs}^{GS} \\ & + \lambda_{i\ell}^{IL} + \lambda_{is}^{IS} + \lambda_{\ell s}^{LS} + \lambda_{g\ell s}^{GLS},\end{aligned}$$

fits well. It is natural to treat injury (*I*) as a response variable and gender (*G*), location (*L*), and seat-belt use (*S*) as explanatory variables, or perhaps *S* as a response with *G* and *L* as explanatory. One can show that this loglinear model is equivalent to logistic model (*G* + *L* + *S*),

$$(9.16) \text{ logit}[P(I = 1|G = g, L = \ell, S = s)] = \alpha + \beta_g^G + \beta_\ell^L + \beta_s^S.$$

For instance, the seat-belt effects in the two models satisfy $\beta_s^S = \lambda_{1s}^{IS} - \lambda_{2s}^{IS}$. In the logit calculation, all terms in the loglinear model not having the injury index *i* cancel. Fitted values, goodness-of-fit statistics, residual df, and standardized residuals for the logistic model are identical to those for the loglinear model.

Odds ratios describing effects on *I* relate to two-factor loglinear parameters and main-effect logistic parameters. In the logistic model, the log odds ratio for the effect of *S* on *I* equals $\beta_1^S - \beta_2^S$. This equals $\lambda_{11}^{IS} + \lambda_{22}^{IS} - \lambda_{12}^{IS} - \lambda_{21}^{IS}$ in the loglinear model. Their estimates are the same no matter how software sets up constraints. For [Table 9.8](#), $\hat{\beta}_1^S - \hat{\beta}_2^S = -0.817$ for the logistic model, and $\hat{\lambda}_{11}^{IS} + \hat{\lambda}_{22}^{IS} - \hat{\lambda}_{12}^{IS} - \hat{\lambda}_{21}^{IS} = -0.817$ for the loglinear model.

Loglinear models are GLMs that treat the 16 cell counts in [Table 9.8](#) as 16 independent Poisson variates. Logistic models are GLMs that treat the table as binomial counts. Logistic models with *I* as the response treat the marginal *GLS* table $\{n_{g+\ell s}\}$ as fixed and regard $\{n_{g\ell s}\}$ as eight independent binomial variates on that response. Although the sampling models differ, the results from fits of corresponding models are identical.

9.5.3 Equivalent Loglinear and Logistic Models

In the derivation of the logistic model $(X + Z)$ [see (9.15)] from loglinear model (XY, XZ, YZ) , the λ^{XZ}_{ik} term cancels. It might seem as if the model (XY, YZ) omitting this term is also equivalent to that logistic model. Indeed, forming the logit on Y for (XY, YZ) results in the same logistic formula. The loglinear model that has the same fit as the logistic model, however, contains a general interaction term for relationships among the explanatory variables. The logistic model does not assume anything about relationships among explanatory variables, so it allows an arbitrary interaction pattern for them.

[Table 9.11](#) summarizes equivalent logistic and loglinear models for three-way tables when Y is a binary response. Each loglinear model contains the XZ association term relating the explanatory variables in the logistic models. The saturated loglinear model (XYZ) is equivalent to a logistic model with an interaction between the predictors X and Z . For instance, the effect of X on Y depends on Z , meaning that the XY odds ratio varies across its categories. That logistic model is also saturated. Analogous correspondences hold when Y has several categories, using baseline-category logit models. An advantage of the loglinear approach is its generality. It applies when more than one response variable exists. The alcohol-cigarette-marijuana example in Section 9.2.4, for instance, used loglinear models to study association patterns among three response variables.

Table 9.11 Equivalent Loglinear and Logistic Models for a Three-Way Table with Binary Response Variable Y

Loglinear Symbol	Logistic Model	Logistic Symbol
(Y, XZ)	α	$(-)$
(XY, XZ)	$\alpha + \beta^X_i$	(X)
(YZ, XZ)	$\alpha + \beta^Z_k$	(Z)
(XY, YZ, XZ)	$\alpha + \beta^X_i + \beta^Z_k$	$(X+Z)$
(XYZ)	$\alpha + \beta^X_i + \beta^Z_k + \beta^{XZ}_{ik}$	$(X*Z)$

9.5.4 Example: Detecting Gene-Environment Interactions in Case-Control Studies

Considerable research in recent years has focused on the role that gene-environment interactions may play in complex diseases. An environmental exposure can markedly increase the risk of a disease in a genetically susceptible subgroup but have little or no effect for others (Umbach and Weinberg 1997). Case-control studies are often used to investigate such interactions.

For a disease outcome Y , binary genetic factor G (such as 1 for the variant genotype and 0 for the “wild type”) and binary environmental factor E , consider the model

$$\text{logit}[P(Y = 1|G = g, E = e)] = \alpha + \beta_1 g + \beta_2 e + \beta_3 ge,$$

where g and e each take values 0 and 1. The corresponding loglinear model is

$$\log \mu_{egy} = \alpha_0 + \beta_0 g + \gamma_0 e + \lambda_0 ge + \alpha y + \beta_1 gy + \beta_2 ey + \beta_3 gey,$$

where α , β_1 , β_2 , and β_3 are the same in each model. Without further restrictions, each model is saturated. For this parameterization, the log odds ratio between G and E is λ_0 when $y = 0$ and $(\lambda_0 + \beta_3)$ when $y = 1$. In case-control studies, Piegorsch et al. (1994) noted that it is often reasonable to assume that $\lambda_0 = 0$, that is, that genotype and environmental exposure are independent in the control population. This assumption corresponds to an unusual instance in which a nonhierarchical model makes sense biologically. The ML estimate of the interaction, $\hat{\beta}_3 = (n_{111}n_{001})/(n_{101}n_{011})$, depends only on the counts for the cases.

Piegorsch et al. (1994) noted that such a case-only analysis can provide a more efficient estimate of the interaction than obtained using all the data. However, results are biased if the independence assumption is violated. Umbach and Weinberg (1997) and Li and Conti (2009) discussed this issue, the latter article using Bayesian model averaging to combine results from case-only and case-control analyses to reduce potential bias. Further complicating issues are that the scale on which gene-environment interaction is most pronounced may not be the one used in standard models, and a statistical interaction may not have a biological interpretation.

9.6 LOGLINEAR MODEL FITTING: LIKELIHOOD EQUATIONS AND ASYMPTOTIC DISTRIBUTIONS

We next discuss loglinear model fitting. After deriving sufficient statistics and likelihood equations, we present large-sample normal distributions for ML estimators of model parameters and cell probabilities. We illustrate results with models for three-way tables. For simplicity, derivations use the Poisson sampling model, which does not require the constraint on $\{\mu_{ijk}\}$ that the multinomial has.

9.6.1 Minimal Sufficient Statistics

For three-way tables, the joint Poisson probability that cell counts $\{Y_{ijk} = n_{ijk}\}$ is

$$\prod_i \prod_j \prod_k \frac{e^{-\mu_{ijk}} \mu_{ijk}^{n_{ijk}}}{n_{ijk}!},$$

with product taken over all cells of the table. The kernel of the log likelihood is

$$(9.17) \quad L(\boldsymbol{\mu}) = \sum_i \sum_j \sum_k n_{ijk} \log \mu_{ijk} - \sum_i \sum_j \sum_k \mu_{ijk}.$$

For the general loglinear model (9.12), this simplifies to

$$(9.18) \quad \begin{aligned} L(\boldsymbol{\mu}) = & n\lambda + \sum_i n_{i++} \lambda_i^X + \sum_j n_{+j+} \lambda_j^Y + \sum_k n_{++k} \lambda_k^Z \\ & + \sum_i \sum_j n_{ij+} \lambda_{ij}^{XY} + \sum_i \sum_k n_{i+k} \lambda_{ik}^{XZ} + \sum_j \sum_k n_{+jk} \lambda_{jk}^{YZ} \\ & + \sum_i \sum_j \sum_k n_{ijk} \lambda_{ijk}^{XYZ} - \sum_i \sum_j \sum_k \exp(\lambda + \dots + \lambda_{ijk}^{XYZ}). \end{aligned}$$

Since the Poisson distribution is in the exponential family, coefficients of the parameters are sufficient statistics. For this saturated model, $\{n_{ijk}\}$ are coefficients of $\{\lambda_{ijk}^{XYZ}\}$, so there is no reduction of the data. For simpler models, certain parameters are zero and (9.18) simplifies. For instance, for the model (X, Y, Z) of mutual independence, sufficient statistics are the coefficients in (9.18) of $\{\lambda_i^X\}$, $\{\lambda_j^Y\}$, and $\{\lambda_k^Z\}$. These are $\{n_{i++}\}$, $\{n_{+j+}\}$, and $\{n_{++k}\}$.

[Table 9.12](#) lists minimal sufficient statistics for several loglinear models. Each one is the coefficient of the highest-order term(s) in which a variable appears. We see that they are the marginal distributions for terms in the model symbol. Simpler models use more condensed sample information. For instance, whereas (X, Y, Z) uses only the single-factor marginal distributions, (XY, XZ, YZ) uses the two-way marginal tables.

Table 9.12 Minimal Sufficient Statistics for Loglinear Models in Three-Way Tables

Model	Minimal Sufficient Statistics
(X, Y, Z)	$\{n_{i++}\}, \{n_{+j+}\}, \{n_{++k}\}$
(XY, Z)	$\{n_{ij+}\}, \{n_{++k}\}$
(XY, YZ)	$\{n_{ij+}\}, \{n_{+jk}\}$
(XY, XZ, YZ)	$\{n_{ij+}\}, \{n_{i+k}\}, \{n_{+jk}\}$

9.6.2 Likelihood Equations for Loglinear Models

The fitted values for a model are solutions to the likelihood equations. We derive likelihood equations in terms of a general formula for a loglinear model. Let $\mathbf{n} = (n_1, \dots, n_N)^T$ and $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)^T$ denote column vectors of observed and expected counts for the N cells of a contingency table, with $n = \sum_i n_i$. For simplicity we use a single index, but the table may be multidimensional. Loglinear models for positive Poisson means have the form

$$(9.19) \log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$$

for a model matrix \mathbf{X} and column vector $\boldsymbol{\beta}$ of model parameters. For example, consider the independence model, $\log \mu_{ij} = \lambda + \lambda^X_i + \lambda^Y_j$, for a 2×2 table. With constraints $\lambda^X_2 = \lambda^Y_2 = 0$, it is

$$\begin{bmatrix} \log \mu_{11} \\ \log \mu_{12} \\ \log \mu_{21} \\ \log \mu_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \lambda^X_1 \\ \lambda^Y_1 \end{bmatrix}.$$

For the model $\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, we have $\log(\mu_i) = \sum_j x_{ij}\beta_j$ for all i . Extending (9.17), for Poisson sampling the log likelihood is

$$(9.20) \quad \begin{aligned} L(\boldsymbol{\mu}) &= \sum_i n_i \log \mu_i - \sum_i \mu_i \\ &= \sum_i n_i \left(\sum_j x_{ij}\beta_j \right) - \sum_i \exp \left(\sum_j x_{ij}\beta_j \right). \end{aligned}$$

The sufficient statistic for β_j is its coefficient, $\sum_i n_i x_{ij}$. Since

$$\begin{aligned} \frac{\partial}{\partial \beta_j} \left[\exp \left(\sum_j x_{ij}\beta_j \right) \right] &= x_{ij} \exp \left(\sum_j x_{ij}\beta_j \right) = x_{ij} \mu_i, \\ \frac{\partial L(\boldsymbol{\mu})}{\partial \beta_j} &= \sum_i n_i x_{ij} - \sum_i \mu_i x_{ij}, \quad j = 1, 2, \dots, p. \end{aligned}$$

The likelihood equations equate these derivatives to zero. They have the form

$$(9.21) \mathbf{X}^T \mathbf{n} = \mathbf{X}^T \hat{\boldsymbol{\mu}}.$$

These equations equate the sufficient statistics to their expected values, a result obtained with GLM theory in (4.32). For models considered so far, these sufficient statistics are the marginal tables in the model symbol. To illustrate, consider model (XZ, YZ) . Its log likelihood is (9.18) with $\lambda^{XY} = \lambda^{XYZ} = 0$. The log-likelihood derivatives

$$\frac{\partial L}{\partial \lambda_{ik}^{XZ}} = n_{i+k} - \mu_{i+k} \quad \text{and} \quad \frac{\partial L}{\partial \lambda_{jk}^{YZ}} = n_{+jk} - \mu_{+jk}$$

yield the likelihood equations

$$(9.22) \hat{\mu}_{i+k} = n_{i+k} \quad \text{for all } i \text{ and } k,$$

$$(9.23) \hat{\mu}_{+jk} = n_{+jk} \quad \text{for all } j \text{ and } k.$$

Derivatives with respect to lower-order terms yield equations implied by these (Exercise 9.29). For model (XZ, YZ) , the fitted values have the same XZ and YZ marginal totals as the observed data.

9.6.3 Unique ML Estimates Match Data in Sufficient Marginal Tables

For model (XZ, YZ), from (9.22), (9.23), and Table 9.12, the minimal sufficient statistics are the ML estimates of the corresponding marginal distributions of expected frequencies. Equation (9.21) gives the corresponding result for any loglinear model. Birch (1963) showed that likelihood equations for loglinear models match minimal sufficient statistics to their expected values. Poisson GLM theory implied this result in (4.32) and (4.51). Thus, fitted values for loglinear models are smoothed versions of the cell counts that match them in certain marginal distributions but have associations and interactions satisfying the model-implied patterns.

Birch showed that a unique set of fitted values both satisfy the model and match the data in the minimal sufficient statistics. Hence, if we find such a solution, it must be the ML solution. To illustrate, the independence model for a two-way table

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$$

has minimal sufficient statistics $\{n_{i+}\}$ and $\{n_{+j}\}$. The likelihood equations are

$$\hat{\mu}_{i+} = n_{i+}, \quad \hat{\mu}_{+j} = n_{+j}, \quad \text{for all } i \text{ and } j.$$

The fitted values [$\hat{\mu}_{ij} = n_{i+} n_{+j}/n$] satisfy these equations and also satisfy the model. Birch's result implies that they are the ML estimates.

9.6.4 Direct Versus Iterative Calculation of Fitted Values

To illustrate how to solve likelihood equations, we continue the analysis of model (XZ, YZ) . From (9.9), the model satisfies

$$\pi_{ijk} = \frac{\pi_{i+k} \pi_{+jk}}{\pi_{++k}} \quad \text{for all } i, j, \text{ and } k.$$

For Poisson sampling, the related formula uses expected frequencies. Setting $\pi_{ijk} = \mu_{ijk}/n$, this is $\{\mu_{ijk} = \mu_{i+k} \mu_{+jk} / \mu_{++k}\}$. The likelihood equations (9.22) and (9.23) specify that ML estimates satisfy $\hat{\mu}_{i+k} = n_{i+k}$ and $\hat{\mu}_{+jk} = n_{+jk}$ and thus also $\hat{\mu}_{++k} = n_{++k}$. Since ML estimates of functions of parameters are the same functions of the ML estimates of those parameters,

$$\hat{\mu}_{ijk} = \frac{\hat{\mu}_{i+k} \hat{\mu}_{+jk}}{\hat{\mu}_{++k}} = \frac{n_{i+k} n_{+jk}}{n_{++k}}.$$

This solution satisfies the model and matches the data in the sufficient statistics. Thus, it is the unique ML solution.

Similar reasoning produces $\{\hat{\mu}_{ijk}\}$ for all except one model in Table 9.12. Table 9.13 shows formulas. That table also expresses $\{\pi_{ijk}\}$ in terms of marginal probabilities. These expressions and the likelihood equations determine the ML formulas, using the approach just described.

Table 9.13 Fitted Values for Loglinear Models in Three-Way Tables

Model ^a	Probabilistic Form	Fitted Value
(X, Y, Z)	$\pi_{ijk} = \pi_{i++} \pi_{+j+} \pi_{++k}$	$\hat{\mu}_{ijk} = \frac{n_{i++} n_{+j+} n_{++k}}{n^2}$
(XY, Z)	$\pi_{ijk} = \pi_{ij+} \pi_{+++}$	$\hat{\mu}_{ijk} = \frac{n_{ij+} n_{+++}}{n}$
(XY, XZ)	$\pi_{ijk} = \frac{\pi_{ij+} \pi_{i+k}}{\pi_{i++}}$	$\hat{\mu}_{ijk} = \frac{n_{ij+} n_{i+k}}{n_{i++}}$
(XY, XZ, YZ)	$\pi_{ijk} = \psi_{ij} \phi_{jk} \omega_{ik}$	Iterative methods (Section 9.7)
(XYZ)	No restriction	$\hat{\mu}_{ijk} = n_{ijk}$

^aFormulas for models not listed are obtained by symmetry; for example, for (XZ, Y) , $\hat{\mu}_{ijk} = n_{i+k} n_{+j+} / n$.

For models having explicit formulas for $\hat{\mu}_{ijk}$, the estimates are said to be *direct*. Many loglinear models do not have direct estimates. ML estimation then requires iterative methods. Of models in Tables 9.12 and 9.13, the only one not having direct estimates is (XY, XZ, YZ) . Although the two-way marginal tables are its minimal sufficient statistics, it is not possible to express $\{\pi_{ijk}\}$ directly in terms of $\{\pi_{ij+}\}$, $\{\pi_{i+k}\}$, and $\{\pi_{+jk}\}$. Direct estimates do not exist for unsaturated models containing all two-factor associations.

9.6.5 Decomposable Models

Unsaturated loglinear models that have direct ML estimates have interpretations in terms of independence, conditional independence, or equiprobability. For those models, expected frequencies decompose into products and ratios of expected marginal sufficient statistics. Such models are called *decomposable* (Andersen 1974). For model (XZ, YZ) of XY conditional independence, for example, from (9.9), $\mu_{ijk} = \mu_{i+k} \mu_{+jk} / \mu_{++k}$.

In practice, it is not essential to know which models have direct estimates. Iterative methods for models not having direct estimates also apply with models that have direct estimates. Statistical software for loglinear models uses such iterative methods for *all* cases.

9.6.6 Chi-Squared Goodness-of-Fit Tests

Model goodness-of-fit statistics compare sample cell counts to fitted counts. For Poisson GLMs, in Section 4.5.2 we showed that for models with an intercept term, the deviance equals the G^2 statistic. With a fixed number of cells, G^2 and X^2 have approximate chi-squared null distributions when expected frequencies are large. The df equal the difference in dimension between the alternative and null hypotheses. This equals the difference between the number of parameters in the general case and when the model holds.

We illustrate with model (X, Y, Z) , for multinomial sampling with probabilities $\{\pi_{ijk}\}$. In the general case, the only constraint is $\sum_i \sum_j \sum_k \pi_{ijk} = 1$, so there are $IJK - 1$ parameters. For model (X, Y, Z) , $\{\pi_{ijk} = \pi_{i++} \pi_{+j+} \pi_{++k}\}$ are determined by $I - 1$ of $\{\pi_{i++}\}$ (since $\sum_i \pi_{i++} = 1$), $J - 1$ of $\{\pi_{+j+}\}$, and $K - 1$ of $\{\pi_{++k}\}$. Thus,

$$df = (IJK - 1) - [(I - 1) + (J - 1) + (K - 1)] = IJK - I - J - K + 2.$$

The same df formula applies for Poisson sampling. Then, the general case has IJK $\{\mu_{ijk}\}$ parameters. The residual df equal the number of cells in the table minus the number of parameters in the Poisson loglinear model for $\{\mu_{ijk}\}$. For instance, model (X, Y, Z) has residual $df = IJK - [1 + (I - 1) + (J - 1) + (K - 1)]$, reflecting the intercept parameter λ and constraints such as $\lambda^X_I = \lambda^Y_J = \lambda^K_K = 0$. This equals the number of linearly independent parameters equated to zero in the saturated model to obtain the given model. [Table 9.14](#) shows df formulas for testing three-way loglinear models.

Table 9.14 Residual Degrees of Freedom for Loglinear Models for Three-Way Tables

Model	Degrees of Freedom
(X, Y, Z)	$IJK - I - J - K + 2$
(XY, Z)	$(K - 1)(IJ - 1)$
(XZ, Y)	$(J - 1)(IK - 1)$
(YZ, X)	$(I - 1)(JK - 1)$
(XY, YZ)	$J(I - 1)(K - 1)$
(XZ, YZ)	$K(I - 1)(J - 1)$
(XY, XZ, YZ)	$I(J - 1)(K - 1)$
(XYZ)	0

9.6.7 Covariance Matrix of ML Parameter Estimators

To present large-sample distributions of ML parameter estimators, we return to general expression $\log(\mu_i) = \sum_j x_{ij}\beta_j$, from which we obtained the log-likelihood derivatives

$$\frac{\partial L(\boldsymbol{\mu})}{\partial \beta_j} = \sum_i n_i x_{ij} - \sum_i \mu_i x_{ij}, \quad j = 1, 2, \dots, p.$$

The Hessian matrix of second partial derivatives has elements

$$\begin{aligned} \frac{\partial^2 L(\boldsymbol{\mu})}{\partial \beta_j \partial \beta_k} &= - \sum_i x_{ij} \frac{\partial \mu_i}{\partial \beta_k} \\ &= - \sum_i x_{ij} \left\{ \frac{\partial}{\partial \beta_k} \left[\exp \left(\sum_h x_{ih} \beta_h \right) \right] \right\} = - \sum_i x_{ij} x_{ik} \mu_i. \end{aligned}$$

Like logistic regression models, loglinear models are GLMs using the canonical link function; thus, this matrix does not depend on the observed data. The information matrix, the negative of this matrix, is

$$\mathcal{J} = X^T \text{Diag}(\boldsymbol{\mu}) X,$$

where $\text{Diag}(\boldsymbol{\mu})$ is a diagonal matrix with the elements of $\boldsymbol{\mu}$ on the main diagonal.

For a fixed number of cells, as $n \rightarrow \infty$, the ML estimator $\hat{\beta}$ is asymptotically normal with mean β and covariance matrix \mathcal{J}^{-1} . Thus, for Poisson sampling, the asymptotic covariance matrix

$$(9.24) \quad \text{cov}(\hat{\beta}) = [X^T \text{Diag}(\boldsymbol{\mu}) X]^{-1}.$$

Substituting ML fitted values and then taking square roots of diagonal elements yields standard errors for $\hat{\beta}$. This also follows from the general expression (4.31) for GLMs, as noted in Section 4.4.9.

9.6.8 Connection Between Multinomial and Poisson Loglinear Models

Similar asymptotic results hold with multinomial sampling. When $\{Y_i, i = 1, \dots, N\}$ are independent Poisson random variables, the conditional distribution of $\{Y_i\}$ given $n = \sum_i Y_i$ is multinomial with parameters $\{\pi_i = \mu_i / (\sum_a \mu_a)\}$. Birch (1963) showed that ML estimates of loglinear model parameters are the same for multinomial sampling as for independent Poisson sampling. He showed that estimates are also the same for independent multinomial sampling, as long as the model contains a term for the marginal distribution fixed by the sampling design. To illustrate, suppose that at each combination of categories of X and Z , an independent multinomial sample occurs on Y . Then, $\{n_{i+k}\}$ are fixed. The model must contain λ^{XZ}_{ik} , so the fitted values satisfy $\{\hat{\mu}_{i+k} = n_{i+k}\}$.

That separate inferential theory is unnecessary for multinomial loglinear models follows from the following argument. Express the Poisson loglinear model for $\{\mu_i\}$ as

$$\log \mu_i = \lambda + \mathbf{x}_i \boldsymbol{\beta},$$

where $(1, \mathbf{x}_i)$ is row i of the model matrix \mathbf{X} and $(\lambda, \boldsymbol{\beta}^T)^T$ is the model parameter vector. The Poisson log likelihood is

$$\begin{aligned} L = L(\lambda, \boldsymbol{\beta}) &= \sum_i n_i \log \mu_i - \sum_i \mu_i \\ &= \sum_i n_i (\lambda + \mathbf{x}_i \boldsymbol{\beta}) - \sum_i \exp(\lambda + \mathbf{x}_i \boldsymbol{\beta}) = n\lambda + \sum_i n_i \mathbf{x}_i \boldsymbol{\beta} - \tau, \end{aligned}$$

where $\tau = \sum_i \mu_i = \sum_i \exp(\lambda + \mathbf{x}_i \boldsymbol{\beta})$. Since $\log \tau = \lambda + \log[\sum_i \exp(\mathbf{x}_i \boldsymbol{\beta})]$, this log likelihood has the form

$$(9.25) \quad L = L(\tau, \boldsymbol{\beta}) = \left\{ \sum_i n_i \mathbf{x}_i \boldsymbol{\beta} - n \log \left[\sum_i \exp(\mathbf{x}_i \boldsymbol{\beta}) \right] \right\} + (n \log \tau - \tau).$$

Now $\pi_i = \mu_i / (\sum_a \mu_a) = \exp(\lambda + \mathbf{x}_i \boldsymbol{\beta}) / [\sum_a \exp(\lambda + \mathbf{x}_a \boldsymbol{\beta})]$, and $\exp(\lambda)$ cancels in the numerator and denominator. Thus, the first term (in braces) on the right-hand side in (9.25) is $\sum_i n_i \log \pi_i$, which is the multinomial log likelihood, conditional on the total cell count n . Unconditionally, $n = \sum_i n_i$ has a Poisson distribution with expectation $\sum_i \mu_i = \tau$, so the second term in (9.25) is the Poisson log likelihood for n . Since $\boldsymbol{\beta}$ enters only in the first term, the ML estimator $\hat{\boldsymbol{\beta}}$ and its covariance matrix for the Poisson log likelihood $L(\lambda, \boldsymbol{\beta})$ are identical to those for the multinomial log likelihood. The Poisson loglinear model has one more parameter (i.e., λ) than the multinomial loglinear model because of the random sample size.

For a multinomial sample, we show in Section 16.4.1 that the estimated covariance matrix of loglinear parameter estimators is

$$(9.26) \quad \widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = \{ \mathbf{X}^T [\text{Diag}(\hat{\boldsymbol{\mu}}) - \hat{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}}^T / n] \mathbf{X} \}^{-1}.$$

The intercept λ from the Poisson model is not relevant, and \mathbf{X} for the multinomial model deletes the column of \mathbf{X} pertaining to it in the Poisson model.

A similar argument applies with several independent multinomial samples. Each log-likelihood term is a sum of components from different samples, but the Poisson log likelihood again decomposes into two parts. One part is a Poisson log likelihood for the independent sample sizes, and the other part is the sum of the independent multinomial log likelihoods. Palmgren (1981) showed that conditional on observed marginal totals for explanatory variables, the asymptotic covariances for estimators of parameters involving the response are the same as for Poisson sampling. For a single multinomial sample, Palmgren's result implies that (9.26) is identical to (9.24) with the row and column referring to λ deleted. Birch (1963), Goodman (1970), and McCullagh and Nelder (1989, p. 211) gave related results. Lang (1996c) gave an elegant discussion of connections between multinomial and Poisson models. His results imply that the asymptotic variance of any linear contrast of estimated log means within a covariate level is identical for the two models.

9.6.9 Distribution of Probability Estimators

For multinomial sampling, the ML estimates of cell probabilities are $\hat{\pi} = \hat{\mu}/n$. We next give the asymptotic $\text{cov}(\hat{\pi})$. Lang (1996c) showed the asymptotic covariance matrix for $\hat{\mu}$ for Poisson sampling and its connection with $\text{cov}(\hat{\pi})$.

The saturated model has $\hat{\pi} = p$, the sample proportions. Under multinomial sampling, from (3.7) and (3.8), their covariance matrix is

$$(9.27) \quad \text{cov}(p) = [\text{Diag}(\pi) - \pi\pi^T]/n.$$

With I independent multinomial samples on a response variable with J categories, π and p consist of I sets of proportions, each having $J - 1$ nonredundant elements. Then, $\text{cov}(p)$ is a block diagonal matrix. Each of the independent samples has a $(J - 1) \times (J - 1)$ block of form (9.27), and the matrix contains zeros off the main diagonal of blocks.

Now assume an unsaturated model. Using the delta method we show in Sections 16.2.2 and 16.4.1 that $\hat{\pi}$ has a large-sample normal distribution about π . The estimated covariance matrix equals

$$\widehat{\text{cov}}(\hat{\pi}) = \widehat{\text{cov}}(p)X[X^T\widehat{\text{cov}}(p)X]^{-1}X^T\widehat{\text{cov}}(p).$$

For tables with many cells, it is not unusual to have a sample proportion of 0 in a cell. In this case the ordinary standard error is 0, which is unappealing. An advantage of fitting a model is that it typically has a positive fitted probability and standard error.

9.6.10 Proof of Uniqueness of ML Estimates

When all $\{n_i > 0\}$, the ML estimates exist and are unique. To show this, for simplicity we use Poisson sampling. Suppose that the model is parameterized so that X has full rank. Birch (1963) showed that the likelihood equations are soluble, by noting that the kernel of the Poisson log likelihood

$$L(\boldsymbol{\mu}) = \sum_i (n_i \log \mu_i - \mu_i)$$

has individual terms converging to $-\infty$ as $\log(\mu_i) \rightarrow \pm\infty$ thus, the log likelihood is bounded above and attains its maximum at finite values of the model parameters. It is stationary at this maximum, since it has continuous first partial derivatives.

Birch showed that the likelihood equations have a unique solution, and the likelihood is maximized at that point. He proved this by showing that the matrix of values $\{-\partial^2 L / \partial \beta_h \partial \beta_j\}$ [i.e., the information matrix $X^T \text{Diag}(\boldsymbol{\mu}) X$] is nonsingular and nonnegative definite, and hence positive definite. Nonsingularity follows from X having full rank and the diagonal matrix having positive elements $\{\mu_i\}$. Any quadratic form $c^T X^T \text{Diag}(\boldsymbol{\mu}) X c$ equals $\sum_i [\sqrt{\mu_i} (\sum_j x_{ij} c_j)]^2 \geq 0$, so the matrix is also nonnegative definite.

9.6.11 Pseudo ML for Complex Sampling Designs

Many surveys have sampling designs employing stratification and/or clustering and have multiple stages. Skinner and Vallet (2010) noted that the meaning of the parameters in the ordinary loglinear model $\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ then depends on the sampling design. It is more sensible to define the model in terms of *population* (rather than sample) expected frequencies $\boldsymbol{\mu}$.

Consider a sampling scheme in which each subject in the population who is classified in cell i is included in the sample with known probability π_i , and let $\boldsymbol{\pi}$ denote the vector of their values. Let $\boldsymbol{\mu}_s$ denote the expected frequencies for the sample of size n . These relate to the population expected frequencies by $\mu_{si} = \pi_i \mu_i$, so they satisfy

$$\log(\boldsymbol{\mu}_s) = \log(\boldsymbol{\pi}) + \mathbf{X}\boldsymbol{\beta}.$$

That is,

$$\log(\mu_{si}) - \log(\pi_i) = \sum_j x_{ij}\beta_j.$$

As noted in Section 4.3.5, the adjustment term, $-\log(\pi_i)$, to the log link is called an *offset*. (In Section 9.7.4 we discuss further the use of model offsets.)

In most sampling designs, however, inclusion probabilities are not constant within cells but vary among the individual subjects, and also the ordinary Poisson sampling model does not apply. For each observation, a case weight (typically the reciprocal of the inclusion probability) inflates or deflates the observation's influence according to features of that design. Adding the case weights for subjects in a particular cell i provides a total weighted frequency for that cell. Denote this by a_i . Skinner and Vallet then denoted the *pseudo ML* estimate of $\boldsymbol{\beta}$ as the solution of the equations

$$\mathbf{X}^T \hat{\boldsymbol{\beta}} = \mathbf{X}^T \hat{\boldsymbol{\mu}},$$

where $\mu_i = \exp(\sum_j x_{ij}\beta_j)$. It can be found using a standard ML fitting routine such as described in the next section, treating the weighted frequencies as the data.

Skinner and Vallet showed that the estimated covariance matrix (9.24) for ordinary Poisson sampling should be replaced by a matrix estimated by

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = [\mathbf{X}^T \text{Diag}(\hat{\boldsymbol{\mu}})\mathbf{X}]^{-1} [\mathbf{X}^T \mathbf{V} \mathbf{X}] [\mathbf{X}^T \text{Diag}(\hat{\boldsymbol{\mu}})\mathbf{X}]^{-1},$$

where \mathbf{V} is an estimator of the covariance matrix of $\boldsymbol{\alpha}$ that accounts for the complex sampling. The matrix \mathbf{V} can be obtained using survey software.² The vector $\boldsymbol{\alpha}$ can be scaled by a constant so the a_i sum to any total, such as the overall sample size n , in which case $[\mathbf{X}^T \text{Diag}(\hat{\boldsymbol{\mu}})\mathbf{X}]^{-1}$ represents the covariance matrix when we ignore the complex sampling design. Skinner and Vallet (2010) showed alternative methods and gave related references.

9.7 LOGLINEAR MODEL FITTING: ITERATIVE METHODS AND THEIR APPLICATION

When a loglinear model does not have direct estimates, iterative algorithms such as Newton–Raphson can solve the likelihood equations. In this section we also present a simpler but more limited method, *iterative proportional fitting*.

9.7.1 Newton–Raphson Method

For the Newton–Raphson method (Section 4.6.1), we identify $L(\beta)$ as the log likelihood for Poisson loglinear models. From (9.20), let

$$L(\beta) = \sum_i n_i \left(\sum_h x_{ih} \beta_h \right) - \sum_i \exp \left(\sum_h x_{ih} \beta_h \right).$$

Then

$$u_j = \frac{\partial L(\beta)}{\partial \beta_j} = \sum_i n_i x_{ij} - \sum_i \mu_i x_{ij},$$

$$h_{jk} = \frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_k} = - \sum_i \mu_i x_{ij} x_{ik},$$

so that

$$u_j^{(t)} = \sum_i (n_i - \mu_i^{(t)}) x_{ij} \quad \text{and} \quad h_{jk}^{(t)} = - \sum_i \mu_i^{(t)} x_{ij} x_{ik}.$$

The t th approximation $\mu^{(t)}$ for $\hat{\mu}$ derives from $\beta^{(t)}$ through $\mu^{(t)} = \exp(X\beta^{(t)})$. It generates the next value $\beta^{(t+1)}$ using (4.45), which in this context is

$$\beta^{(t+1)} = \beta^{(t)} + [X^T \text{Diag}(\mu^{(t)}) X]^{-1} X^T (\mathbf{n} - \mu^{(t)}).$$

This in turn produces $\mu^{(t+1)}$, and so on.

Alternatively, $\beta^{(t+1)}$ can be expressed as

$$(9.28) \quad \beta^{(t+1)} = -(\mathbf{H}^{(t)})^{-1} \mathbf{r}^{(t)},$$

where $r_j^{(t)} = \sum_i \mu_i^{(t)} x_{ij} [\log \mu_i^{(t)} + (n_i - \mu_i^{(t)})/\mu_i^{(t)}]$. The expression in brackets is the first term in the Taylor series expansion of $\log n_i$ at $\log \mu_i^{(t)}$.

The iterative process begins with all $\mu_i^{(0)} = n_i$, or with an adjustment such as $\mu_i^{(0)} = n_i + \frac{1}{2}$ if any $n_i = 0$. Then (9.28) produces $\beta^{(1)}$, and for $t > 0$ the iterations proceed as just described with $\{n_i\}$. For loglinear models $L(\beta)$ is concave, and $\mu^{(t)}$ and $\beta^{(t)}$ usually converge rapidly to the ML estimates $\hat{\mu}$ and $\hat{\beta}$ as t increases. The $\mathbf{H}^{(t)}$ matrix converges to $\hat{\mathbf{H}} = X^T \text{Diag}(\hat{\mu}) X$. By (9.24), the estimated large-sample covariance matrix of $\hat{\beta}$ is $-\hat{\mathbf{H}}^{-1}$, a by-product of the method.

As we discussed in Section 4.6.4 for GLMs, (9.28) has the iterative reweighted least-squares form

$$\beta^{(t+1)} = (X^T \hat{\mathbf{V}}_t^{-1} X)^{-1} X^T \hat{\mathbf{V}}_t^{-1} z^{(t)}.$$

Here, $z^{(t)}$ has elements $n_i = \log \mu_i^{(t)} + (n_i - \mu_i^{(t)})/\mu_i^{(t)}$ and $\hat{\mathbf{V}}_t = [\text{Diag}(\mu^{(t)})]^{-1}$. Thus, $\beta^{(t+1)}$ is the weighted least-squares solution for a model

$$z^{(t)} = X\beta + \epsilon,$$

where $\{\epsilon_i\}$ are uncorrected with variances $\{1/\mu_i(t)\}$. With $\{\mu_i^{(0)} = n_i\}$, $\beta^{(1)}$ is the weighted least-squares estimate for model $\log(\mathbf{n}) = X\beta + \epsilon$.

9.7.2 Iterative Proportional Fitting

The *iterative proportional fitting* (IPF) algorithm is a simple method for calculating $\{\mu_i\}$ for loglinear models. Introduced by Deming and Stephan (1940) and later extended by Bishop (1969) and by Fienberg (1970a) for loglinear modeling, it has the following steps:

1. Start with $\{\mu_i^{(0)}\}$ satisfying a model no more complex than the one being fitted. For instance, $\{\mu_i^{(0)} \equiv 1.0\}$ are trivially adequate.
2. By multiplying by appropriate factors, adjust $\{\mu_i^{(0)}\}$ successively to match each marginal table in the set of minimal sufficient statistics.
3. Continue until the maximum difference between the sufficient statistics and their fitted values is sufficiently close to zero.

We illustrate using model (XY, XZ, YZ) . Its minimal sufficient statistics are $\{n_{ij+}\}$, $\{n_{i+k}\}$, and $\{n_{+jk}\}$. Initial estimates must satisfy the model. The first cycle of the IPF algorithm has three steps:

$$\mu_{ijk}^{(1)} = \mu_{ijk}^{(0)} \frac{n_{ij+}}{\mu_{ij+}^{(0)}}, \quad \mu_{ijk}^{(2)} = \mu_{ijk}^{(1)} \frac{n_{i+k}}{\mu_{i+k}^{(1)}}, \quad \mu_{ijk}^{(3)} = \mu_{ijk}^{(2)} \frac{n_{+jk}}{\mu_{+jk}^{(2)}}.$$

Summing both sides of the first expression over k shows that $\mu_{ij+}^{(1)} = n_{ij+}$ for all i and j . After step 1, observed and fitted values match in the XY marginal table. After step 2, all $\mu_{i+k}^{(2)} = n_{i+k}$, but the XY marginal tables no longer match. After step 3, all $\mu_{+jk}^{(3)} = n_{+jk}$, but the XY and XZ marginal tables no longer match. A new cycle begins by again matching the XY marginal tables, using $\mu_{ijk}^{(4)} = \mu_{ijk}^{(3)}(n_{ij+}/\mu_{ij+}^{(3)})$, and so on.

At each step, the updated estimates continue to satisfy the model. For instance, step 1 uses the same adjustment factor $(n_{ij+}/\mu_{ij+}^{(0)})$ at different levels k of Z . Thus, XY odds ratios from different levels of Z have ratio equal to 1, and the homogeneous association pattern continues at each step.

As the cycles progress, the G^2 statistic comparing cell counts to the updated fit is monotone decreasing, and the process must converge (Fienberg 1970a, Haberman 1974a). The IPF algorithm produces ML estimates because it generates a sequence of fitted values converging to a solution that both satisfies the model and matches the sufficient statistics. From Section 9.6.10, only one such solution exists, and it is ML.

The IPF method works even for models having direct estimates. Then, IPF normally yields ML estimates within one cycle (Haberman 1974a, p. 197). We illustrate with the model of independence. The minimal sufficient statistics are $\{n_{i+}\}$ and $\{n_{+j}\}$. With $\{\mu_{ij}^{(0)} \equiv 1.0\}$, the first cycle gives

$$\begin{aligned} \mu_{ij}^{(1)} &= \mu_{ij}^{(0)} \frac{n_{i+}}{\mu_{i+}^{(0)}} = \frac{n_{i+}}{J}, \\ \mu_{ij}^{(2)} &= \mu_{ij}^{(1)} \frac{n_{+j}}{\mu_{+j}^{(1)}} = \frac{n_{i+}n_{+j}}{n}. \end{aligned}$$

The IPF algorithm then gives $\hat{\mu}_{ij}^{(t)} = n_{i+}n_{+j}/n$ for all $t > 2$.

9.7.3 Comparison of IPF and Newton–Raphson Iterative Methods

The IPF algorithm is simple and easy to implement. It converges to the ML fit even when the likelihood is poorly behaved, for instance, with zero fitted counts and estimates on the boundary of the parameter space. The Newton–Raphson method is more complex, requiring solving a system of equations at each step. Newton–Raphson is sometimes not feasible when the model is of high dimensionality—for instance, when the contingency table and parameter vector are huge.

However, IPF has disadvantages. It is applicable primarily to models for which likelihood equations equate observed and fitted counts in marginal tables. By contrast, Newton–Raphson is a general-purpose method that can solve more complex likelihood equations. IPF sometimes converges slowly compared with Newton–Raphson. Unlike Newton–Raphson, IPF does not produce the model parameter estimates and their estimated covariance matrix as a by-product. Fitted values that IPF produces can generate this information. Model parameter estimates are contrasts of $\{\log \hat{\mu}_i\}$ (see Exercises 9.16 and 9.17), and substituting fitted values into (9.24) yields $\text{cov}(\hat{\beta})$.

Because the Newton–Raphson algorithm applies to a wide variety of models and also yields standard errors, it is the fitting routine used by most software for loglinear models. IPF is primarily of historical interest. However, for some applications the analysis is more transparent using IPF. The next example illustrates.

9.7.4 Raking a Table: Contingency Table Standardization

[Table 9.15](#) relates political party affiliation and political ideology for the 2008 General Social Survey. To make the pattern of association clearer, we standardized the table so that all row and column marginal totals equal 100 while maintaining the sample odds ratio structure.

Table 9.15 Marginal Standardization of Political Ideology by Political Party Affiliation

Party Affiliation	Political Ideology			Total
	Liberal	Moderate	Conservative	
Democrat	306 (55.0)	279 (32.5)	116 (12.5)	(100)
Independent	185 (36.7)	312 (40.1)	194 (23.2)	(100)
Republican	26 (8.2)	134 (27.5)	338 (64.3)	(100)
Total	(100)	(100)	(100)	

The IPF routine to standardize with margins of 100 is

$$\mu_{ij}^{(0)} = n_{ij}$$

and then for $t = 1, 3, 5, \dots$,

$$\mu_{ij}^{(t)} = \mu_{ij}^{(t-1)} \frac{100}{\mu_{i+}^{(t-1)}}, \quad \mu_{ij}^{(t+1)} = \mu_{ij}^{(t)} \frac{100}{\mu_{+j}^{(t)}}.$$

At the end of each odd-numbered step, all row totals equal 100. At the end of each even-numbered step, all column totals equal 100. Odds ratios do not change at each odd (even) step, since all counts in a given row (column) multiply by the same constant.

The IPF algorithm converges to the entries in parentheses in [Table 9.15](#). The association is clearer in this standardized table. A ridge appears down the main diagonal, with Republicans having more conservative political ideology. The other counts fall away smoothly on both sides.

Table standardization is a useful method for comparing tables having different marginal structures. Mosteller (1968) compared intergenerational occupational mobility tables from Britain and Denmark. Yule (1912) compared three hospitals on vaccination and recovery for smallpox patients. A modern application is adjusting sample data to match marginal distributions specified by census results.

The process of table standardization is called *raking* the table. Imrey et al. (1981) and Little and Wu (1991) derived the asymptotic covariance matrix for raked sample proportions. For sample counts $\{n_{ij}\}$ with $\{\mu_{ij} = E(n_{ij})\}$, let $\{E_{ij}\}$ denote expected frequencies for the standardized table and $\{\hat{E}_{ij}\}$ fitted values in the standardized table. The standardization process corresponds to fitting the model

$$\log(E_{ij}/\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y.$$

That is, maintaining the odds ratios means that the two-way tables of $\{E_{ij}/\mu_{ij}\}$ and of $\{\hat{E}_{ij}/n_{ij}\}$ satisfy independence.

The fitted values $\{\hat{E}_{ij}\}$ in the standardized table satisfy

$$\log \hat{E}_{ij} - \log n_{ij} = \hat{\lambda} + \hat{\lambda}_i^X + \hat{\lambda}_j^Y,$$

with offset $-\log n_{ij}$. Standard GLM software can fit models having offsets. To rake a table, one enters as sample data pseudo-values that satisfy independence and have the desired margins, taking $\log n_{ij}$ as an offset.

NOTES

Section 9.2: Loglinear Models for Independence and Interaction in Three-Way Tables

9.1 Early loglinear: Roy and Mitra (1956) discussed types of independence for three-way tables and their large-sample tests. Birch's (1963) article on ML estimation for loglinear models was part of substantial research on loglinear models in the 1960s, much due to L. A. Goodman (see Section 17.4). Haberman (1974a) presented an influential theoretical study of loglinear models.

Section 9.3: Inference for Loglinear Models

9.2 Decomposable models: Decomposable models were studied by Andersen (1974) and Sundberg (1975), building on earlier results by Goodman (1970, 1971b) and Haberman (1974a, Chap. 5), who proved conditions under which loglinear models have direct estimates. For later related work, see Darroch et al. (1980), Dobra (2003), Dobra and Fienberg (2000), Lauritzen (1996), and Whittaker (1990, Sec. 12.4). Meeden et al. (1998) showed that with squared error loss, the ML estimator of cell probabilities for any decomposable loglinear model is admissible. Baglivo et al. (1992) and Forster et al. (1996) discussed small-sample exact inference. See also Section 10.1.2.

9.3 Measurement error: For methods that allow for misclassification error, see Espeland and Hui (1987), Kuha and Skinner (1997), Kuha et al. (2005) and references therein, and Palmgren and Ekholm (1987). For the related issue of measurement error, see Buonaccorsi (2010, Chap. 2, 3, 7) and Cox and Snell (1989, Sec. 3.4).

Section 9.7: Loglinear Model Fitting: Iterative Methods and Their Application

9.4 IPF: Darroch (1962) used IPF to obtain ML estimates in contingency tables. Bishop et al. (1975), Fienberg (1970a), and Speed (2005) presented other applications of IPF. Darroch and Ratcliff (1972) generalized IPF for models in which sufficient statistics are more complex than marginal tables.

9.5 Raking: For further discussion of table raking, see Bishop et al. (1975, pp. 99–102), Deville et al. (1993), Haberman (1979, Chap. 9), Hoem (1987), Imrey et al. (1981), and Little and Wu (1991).

EXERCISES

Applications

9.1 A General Social Survey asked: “Do you support or oppose the following measures to deal with AIDS? (1) Have the government pay all of the health care costs of AIDS patients; (2) Develop a government information program to promote safe sex practices, such as the use of condoms.” [Table 9.16](#) summarizes opinions about health care costs (H) and the information program (I), classified also by the respondent’s gender (G). Fit loglinear models (GH, GI) , (GH, HI) , (GI, HI) , and (GH, GI, HI) . Show that models that lack the HI term fit poorly. Interpret results for the model (GH, GI, HI) .

[Table 9.16](#) Data for Exercise 9.1 on Measures for Dealing with AIDS

Gender	Information Opinion	Health Opinion	
		Support	Oppose
Male	Support	76	160
	Oppose	6	25
Female	Support	114	181
	Oppose	11	48

9.2 [Table 9.17](#) shows the result of cross-classifying a sample of people from the MBTI Step II National Sample, collected and compiled by CPP, Inc., on the four scales of the Myers–Briggs personality test: Extroversion/Introversion (E/I), Sensing/iNtuitive (S/N), Thinking/Feeling (T/F), and Judging/Perceiving (J/P). The 16 cells in this table correspond to the personality types. Fit the loglinear model of homogeneous association. Based on the fit, show that the estimated conditional association is strongest between the S/N and J/P scales and that there is not strong evidence of conditional association between the E/I and T/F scales or between the E/I and J/P scales.

[Table 9.17](#) Data on Four Scales of Myers–Briggs Personality Test

Extroversion/Introversion		E				I			
		S		N		S		N	
				Thinking/Feeling					
Judging/Perceiving		T	F	T	F	T	F	T	F
J		77	106	23	31	140	138	13	31
P		42	79	18	80	52	106	35	79

Source: Reproduced with special permission of CPP, Inc., Mountain View, CA 94043.

9.3 Refer to the previous exercise. [Table 9.18](#) shows the fit of the model that assumes conditional independence between E/I and T/F and between E/I and J/P but has the other pairwise associations.

[Table 9.18](#) Software Output (Based on SAS) for Fitting a Loglinear Model to [Table 9.17](#)

Criteria For Assessing Goodness Of Fit						
Criterion	DF	Value				
Deviance	7	12.3687				
Pearson Chi-Square	7	12.1996				
Analysis Of Parameter Estimates						
Parameter	DF	Estimate	Standard Error	LR 95% Confidence Limits	Wald Chi-Square	
EI*SN	e n	1	0.3219	0.1360 0.0553 0.5886	5.60	
SN*TF	n f	1	0.4237	0.1520 0.1278 0.7242	7.77	
SN*JP	n j	1	-1.2202	0.1451 -1.5075 -0.9382	70.69	
TF*JP	f j	1	-0.5585	0.1350 -0.8242 -0.2948	17.12	

- a. Compare this to the fit of the model containing all the pairwise associations, which has deviance 10.16 with df = 5. What do you conclude?
- b. Show how to use the limits reported to construct a 95% profile likelihood confidence

interval for the conditional odds ratio between the S/N and J/P scales. Interpret.

c. SAS (PROC GENMOD) reports maximized log-likelihood values of 3475.19 for the mutual independence model, 3538.05 for the homogeneous association model, and 3539.58 for the model containing all the three-factor interaction terms. Write the loglinear model for each case, and show that the numbers of model parameters are 5, 11, and 15, so residual df = 11, 5, and 1.

d. According to AIC, which model in (c) seems best? Why?

9.4 Refer to Section 9.3.2. Explain why software for which parameters sum to zero across levels of each index reports $\hat{\lambda}^{AC}_{11} = \hat{\lambda}^{AC}_{22} = 0.514$ and $\hat{\lambda}^{AC}_{12} = \hat{\lambda}^{AC}_{21} = -0.514$, with $SE = 0.044$ for each term.

9.5 Subjects in a GSS were asked their opinions about government spending on the environment (*E*), health (*H*), assistance to big cities (*C*), and law enforcement (*L*). The data are shown at the text website, with outcome categories 1 = too little, 2 = about right, 3 = too much. For the homogeneous association model, [Table 9.19](#) shows some results, including the two-factor estimates for the *EH* association for coding by which estimates at category 3 of each variable equal 0.

Table 9.19 Software Output (Based on SAS) for Fitting Model For Exercise 9.5

Criteria For Assessing Goodness Of Fit							
Criterion	DF	Value	Value/DF				
Deviance	48	31.6695	0.6598				
Pearson Chi-Square	48	26.5224	0.5526				
Log Likelihood		1284.9404					

Parameter	DF	Estimate	Standard	Wald	95%	Chi-	Square
			Error	Confidence	Limits	Square	
e*h	1	1	2.1425	0.5566	1.0515	3.2335	14.81
e*h	1	2	1.4221	0.6034	0.2394	2.6049	5.55
e*h	2	1	0.7294	0.5667	-0.3813	1.8402	1.66
e*h	2	2	0.3183	0.6211	-0.8991	1.5356	0.26

a. Test the model goodness of fit, and interpret.

b. Report the estimated *EH* conditional odds ratio for the (i) too much and too little categories, (ii) too much and about right categories, and (iii) about right and too little categories.

c. [Table 9.20](#) reports $\{\hat{\lambda}_{eh}^{EH}\}$ when parameters sum to zero within rows and within columns, and when parameters are zero in the first row and first column. Show how these yield the estimated *EH* conditional odds ratio for the too much and too little categories. Construct a confidence interval for that odds ratio. Interpret.

Table 9.20 Parameter Estimates for Model in Exercise 9.5

<i>E</i>	Sum to Zero Constraints			Zero for First Level		
	<i>H</i>			<i>H</i>		
	1	2	3	1	2	3
1	0.509	0.166	-0.676	0	0	0
2	-0.065	-0.099	0.163	0	0.309	1.413
3	-0.445	-0.068	0.513	0	0.720	2.142

9.6 For 2010 General Social Survey data cross-classifying opinions on *A* = abortion should be legal for any reason (1 = yes, 0 = no), *E* = willingness to pay higher taxes to help the environment (1 = yes, 0 = no), and *P* = political party identification (1 = Democratic, 0 = Republican), the 8 cell counts for (*A*, *E*, *P*) values were 50 for (1,1,1), 5 for (1,1,0), 28 for (1,0,1), 30 for (1,0,0), 32 for (0,1,1), 27 for (0,1,0), 50 for (0,0,1), 61 for (0,0,0). Analyze these data using loglinear models. (We do not list the counts here for Independents or for those who were neutral about higher taxes for the environment)

9.7 [Table 9.21](#) refers to automobile accident records in Florida.

Table 9.21 Data for Exercise 9.7

Safety Equipment in Use	Whether Ejected	Injury	
		Nonfatal	Fatal
Seat belt	Yes	1,105	14
	No	411,111	483
None	Yes	4,624	497
	No	157,342	1,008

Source: Florida Department of Highway Safety and Motor Vehicles.

- a. Find a loglinear model that describes the data well. Interpret associations.
- b. Treating whether killed as the response, fit an equivalent logistic model. Interpret the effects.
- c. Since n is large, goodness-of-fit statistics are large unless the model fits very well. Calculate the dissimilarity index for the model in part (a), and interpret.

9.8 Refer to the loglinear models for the auto accident data of [Table 9.8](#).

- a. Explain why the fitted odds ratios in [Table 9.10](#) for model (GI, GL, GS, IL, IS, LS) suggest that the most likely accident case for injury is females not wearing seat belts in rural locations.
- b. Fit model (GLS, GI, IL, IS) . Using model parameter estimates, show that the fitted IS conditional odds ratio equals 0.44, and show that for each injury level, the estimated conditional LS odds ratio is 1.17 for ($G = \text{female}$) and 1.03 for ($G = \text{male}$).
- c. Consider the following two-stage model: The first stage is a logistic model with S as the response for the three-way GLS table. The second stage is a logistic model with these three variables as predictors for I in the four-way table. Explain why this composite model is sensible, fit the models, and interpret results.

9.9 Refer to the logistic model in Exercise 5.17 on the death penalty.

- a. Give the symbol for the loglinear model that is equivalent to this logistic model.
- b. Which logistic model corresponds to loglinear model (YD, YV, DVF) ?
- c. State the equivalent loglinear and logit models for which (i) Y is jointly independent of D , V , and F ; (ii) there are main effects of F on Y , but Y is conditionally independent of D and V , given F ; and (iii) there is interaction between D and V in their effects on Y , and F has main effects.

9.10 For a multiway contingency table, when is a logistic model more appropriate than a loglinear model? When is a loglinear model more appropriate?

9.11 Using software, conduct the analyses described in this chapter for the student survey data ([Table 9.3](#)).

9.12 Using table raking, standardize [Table 11.7](#). Describe the migration patterns.

9.13 The book's website (www.stat.ufl.edu/~aa/cda/cda.html) has a $2 \times 3 \times 2 \times 2$ table relating responses on frequency of attending religious services, political views, opinion on making birth control available to teenagers, and opinion about a man and woman having sexual relations before marriage. Analyze these data using loglinear models. Interpret results.

Theory and Methods

9.14 Suppose that $\{\mu_{ij} = n\pi_{ij}\}$ satisfy the independence model [\(9.1\)](#).

- a. Show that $\lambda_a^Y - \lambda_b^Y = \log(\pi_{+a}/\pi_{+b})$.
- b. Show that $\{\text{all } \lambda_j^Y = 0\}$ is equivalent to $\pi_{+j} = 1/J$ for all j .

9.15 Refer to the independence model, $\mu_{ij} = \mu\alpha_i\beta_j$. For the corresponding loglinear model [\(9.1\)](#):

- a. Show that you can constrain $\sum_i \lambda_i^X = \sum_j \lambda_j^Y = 0$ by setting

$$\lambda_i^X = \log \alpha_i - \left(\sum_h \log \alpha_h \right) / I, \quad \lambda_j^Y = \log \beta_j - \left(\sum_h \log \beta_h \right) / J,$$

$$\lambda = \log \mu + \left(\sum_h \log \alpha_h \right) / I + \left(\sum_h \log \beta_h \right) / J.$$

b. Show that you can constrain $\lambda_1^X = \lambda_1^Y = 0$ by defining $\lambda_i^X = \log \alpha_i - \log \alpha_1$ and $\lambda_j^Y = \log \beta_j - \log \beta_1$. Then, what does λ equal?

9.16 For an $I \times J$ table, let $\eta_{ij} = \log \mu_{ij}$, and let a dot subscript denote the mean for that index (e.g., $\eta_{..} = \sum_j \eta_{ij} / J$). Then, let $\lambda = \eta_{..}$, $\lambda_i^X = \eta_{i..} - \eta_{..}$, $\lambda_j^Y = \eta_{.j} - \eta_{..}$, and $\lambda_{ij}^{XY} = \eta_{ij} - \eta_{i..} - \eta_{.j} + \eta_{..}$.

a. Show that $\mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$. Hence, any set of positive $\{\mu_{ij}\}$ satisfies the saturated model.

b. Show that $\sum_i \lambda_i^X = \sum_j \lambda_j^Y = \sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = 0$.

c. For 2×2 tables, show that $\log \theta = 4\lambda_{11}^{XY}$.

d. For $2 \times J$ tables, show that $\lambda_{11}^{XY} = (\sum_j \log \alpha_j) / 2J$, where $\alpha_j = \mu_{11}\mu_{2j} / \mu_{21}\mu_{1j}$, $j = 2, \dots, J$.

e. Alternative constraints have other odds ratio formulas. Let $\lambda = \eta_{11}$, $\lambda_i^X = \eta_{i1} - \eta_{11}$, $\lambda_j^Y = \eta_{1j} - \eta_{11}$, and $\lambda_{ij}^{XY} = \eta_{ij} - \eta_{i1} - \eta_{1j} + \eta_{11}$. Then, show that the saturated model holds with $\lambda_i^X = \lambda_j^Y = \lambda_{1j}^{XY} = \lambda_{i1}^{XY} = 0$ for all i and j , and $\lambda_{ij}^{XY} = \log(\mu_{11}\mu_{ij}/\mu_{1j}\mu_{i1})$.

9.17 Suppose that all $\mu_{ijk} > 0$. Let $\eta_{ijk} = \log \mu_{ijk}$, and consider model parameters with zero-sum constraints.

a. For the general loglinear model (9.12), define parameters in the fashion of Exercise 9.16 (e.g., $\lambda_{ij}^{XY} = \eta_{ij..} - \eta_{i..} - \eta_{.j..} + \eta_{...}$).

b. For model (XY, XZ, YZ) with a $2 \times 2 \times 2$ table, show that $\lambda_{11}^{XY} = \frac{1}{4} \log \theta_{11(1)}$.

c. For (XYZ) with a $2 \times 2 \times 2$ table, show that

$$\lambda_{111}^{XYZ} = \frac{1}{8} \log[\theta_{11(1)} / \theta_{11(2)}].$$

9.18 Two balanced coins are flipped, independently. Let X = whether the first flip resulted in a head (yes, no), Y = whether the second flip resulted in a head, and Z = whether both flips had the same result. Using this example, show that marginal independence for each pair of three variables does not imply that the variables are mutually independent.

9.19 For three categorical variables X , Y , and Z :

a. When Y is jointly independent of X and Z , show that X and Y are conditionally independent, given Z .

b. Prove that mutual independence of X , Y , and Z implies that X and Y are both marginally and conditionally independent.

c. When X is independent of Y and Y is independent of Z , explain why X is not necessarily independent of Z .

9.20 Suppose X and Y are conditionally independent, given Z , and X and Z are marginally independent. Show that X and Y are also marginally independent.

9.21 A $2 \times 2 \times 2$ table satisfies $\pi_{i++} = \pi_{+j+} = \pi_{++k} = \frac{1}{2}$, all i, j, k . Give an example of $\{\pi_{ijk}\}$ that satisfies model **(a)** (X, Y, Z) , **(b)** (XY, Z) , **(c)** (XY, YZ) , **(d)** (XY, XZ, YZ) , and **(e)** (XYZ) , but in each case not a simpler model.

9.22 Suppose model (XY, XZ, YZ) holds in a $2 \times 2 \times 2$ table, and the common XY conditional log odds ratio at the two levels of Z is positive. If the XZ and YZ conditional log odds ratios are both positive or both negative, show that the XY marginal odds ratio is larger than the XY conditional odds ratio. Hence, Simpson's paradox cannot occur for the XY association.

9.23 Show that the general loglinear model in T dimensions has 2^T terms. [Hint: It has an intercept, $\binom{T}{1}$ single-factor terms, $\binom{T}{2}$ two-factor terms,]

9.24 Each of T responses is binary. For indicator variables $\{z_1, \dots, z_T\}$, the loglinear model of

mutual independence has the form

$$\log \mu_{z_1, \dots, z_T} = \lambda_1 z_1 + \dots + \lambda_T z_T.$$

Show how to express the general loglinear model (Cox 1972).

9.25 Consider a cross-classification of W, X, Y, Z .

a. Explain why (WXZ, WYZ) is the most general loglinear model for which X and Y are conditionally independent.

b. State the model symbol for which X and Y are conditionally independent *and* there is no three-factor interaction.

9.26 For a four-way table with binary response Y , give the equivalent loglinear and logistic models that have main effects of factors A, B , and C on Y when (a) Y is binary and (b) Y has $J > 2$ categories.

9.27 For the independence model for a two-way table, derive minimal sufficient statistics, likelihood equations, fitted values, and residual df.

9.28 For the loglinear model for an $I \times J$ table, $\log \mu_{ij} = \lambda + \lambda_i^X$, show that $\hat{\mu}_{ij} = n_{i+}/J$ and residual df = $I(J - 1)$.

9.29 Write the log likelihood L for model (XZ, YZ) . Calculate $\partial L / \partial \lambda$ and show that it implies $\hat{\mu}_{+++} = n$. Show that $\partial L / \partial \lambda_i^X = n_{i++} - \mu_{i++}$. Similarly, differentiate with respect to each parameter to obtain likelihood equations. Show (9.22) and (9.23) imply the other equations, so those equations determine the ML estimates.

9.30 Consider the loglinear model with symbol (XZ, YZ) .

a. For fixed k , show that $\{\hat{\mu}_{ijk}\}$ equal the fitted values for testing independence between X and Y within level k of Z .

b. Show that the Pearson and likelihood-ratio statistics for testing this model's fit have form $X^2 = \sum X_k^2$, where X_k^2 tests independence between X and Y at level k of Z .

9.31 Table 9.22 shows fitted values for models for four-way tables that have direct estimates.

Table 9.22 Fits of Four-Way-Table Loglinear Models for Exercise 9.31^a

Model	Expected Frequency Estimate	Residual DF
(W, X, Y, Z)	$n_{h+++}n_{i++}n_{++j}n_{++k}/n^3$	$HJK - H - I - J - K + 3$
(WX, Y, Z)	$n_{hi++}n_{++j}n_{++k}/n^2$	$HJK - HI - J - K + 2$
(WX, WY, Z)	$n_{hi++}n_{h+j}n_{++k}/n_{h+++}n$	$HJK - HI - HJ - K + H + 1$
(WX, YZ)	$n_{hi++}n_{++jk}/n$	$(HI - 1)(JK - 1)$
(WX, WY, XZ)	$n_{hi++}n_{h+j}n_{i+k}/n_{h+++}n_{i++}$	$HJK - HI - HJ - IK + H + I$
(WX, WY, WZ)	$n_{hi++}n_{h+j}n_{h+k}/(n_{h+++})^2$	$HJK - HI - HJ - HK + 2H$
(WXY, Z)	$n_{hij}n_{++k}/n$	$(HIJ - 1)(K - 1)$
(WXY, WZ)	$n_{hij}n_{h+k}/n_{h++}$	$H(IJ - 1)(K - 1)$
(WXY, WXZ)	$n_{hij}n_{hi+k}/n_{hi++}$	$HI(J - 1)(K - 1)$

^aNumber of levels of W, X, Y, Z , denoted by H, I, J, K . Estimates for other models of each type are obtained by symmetry.

- a. Use Birch's results to verify that the entry is correct for (W, X, Y, Z) . Verify its residual df.
b. Motivate the estimate and df formulas for (WX, YZ) , (WXY, Z) , (WXY, WZ) , and (WXY, WXZ) using composite variables and the corresponding results for two-way tables [e.g., for (WXY, WZ) , given W, Z is independent of the composite XY variable].

9.32 A T -dimensional table $\{n_{ab\dots t}\}$, has I_i categories in dimension i .

a. For the mutual independence model, explain why the minimal sufficient statistics are the one-way marginal distributions, the fitted probabilities are the product of the T one-dimensional marginal proportions, and residual df = $\prod_i I_i - [1 + \sum_i (I_i - 1)] = \prod_i I_i - \sum_i I_i + T - 1$.

b. For the hierarchical homogeneous association model having all two-factor associations but no three-factor interactions, explain why the minimal sufficient statistics are all the two-factor marginal distributions, and the residual df = $\prod_i I_i - [1 + \sum_i (I_i - 1) + \sum \sum_{i < j} (I_i - 1)(I_j - 1)]$.

9.33 Consider loglinear model (X, Y, Z) for a $2 \times 2 \times 2$ table.

- a. Express the model in the form $\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$.
- b. Show that the likelihood equations $\mathbf{X}^T \mathbf{n} = \mathbf{X}^T \hat{\boldsymbol{\mu}}$ equate $\{n_{ijk}\}$ and $\{\hat{\mu}_{ijk}\}$ in the one-dimensional margins.

9.34 Apply IPF to model (a) (X, YZ) and (b) (XZ, YZ) . Show that the ML estimates result within one cycle.

9.35 Refer to Section 9.6.3. Show that L has individual terms converging to $-\infty$ as $\log \mu_i \rightarrow \pm\infty$. Explain why positive definiteness of the information matrix implies that the solution of the likelihood equations is unique, with likelihood maximized at that point.

¹The X and Y superscripts represent the variables and are not exponents.

²Such as Survey Analysis in R.

CHAPTER 10

Building and Extending Loglinear Models

From Chapter 9, loglinear models for contingency tables use the log link for Poisson cell counts in describing associations and interactions among a set of categorical response variables, and connections exist between them and logistic models. In this chapter we discuss topics dealing with building and extending loglinear models.

In Section 10.1 we show how certain models having a conditional independence structure can be represented by graphs. In Section 10.2 we discuss selection and comparison of loglinear models. Diagnostics for checking models, such as residuals, are presented in Section 10.3. The loglinear models of Chapter 9 treat all variables as nominal. Generalizations of loglinear models and related *association models* and *correlation models* can also describe association between ordinal variables. One approach scores those variables with fixed or parameter scores, as Sections 10.4 and 10.5 show. In Section 10.6 we cover complications that can occur with sparse contingency tables. In the final section we discuss Bayesian loglinear modeling.

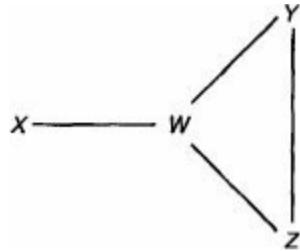
10.1 CONDITIONAL INDEPENDENCE GRAPHS AND COLLAPSIBILITY

Many loglinear models can be portrayed with a graph that represents the conditional independence structure among the responses. This representation also helps to reveal implications of models, such as when an association is unchanged when we collapse a table over another variable.

10.1.1 Conditional Independence Graphs

A *graph*, according to graph theory, consists of two sets: a set of vertices and a set of edges connecting some vertices. In a *conditional independence graph*, each vertex represents a variable. The absence of an edge connecting two variables represents a conditional independence between them. For instance, loglinear model (WX, WY, WZ, YZ) lacks XY and XZ terms. It assumes independence between X and Y and between X and Z , conditional on the remaining two variables. [Figure 10.1](#) portrays this model's graph. The four variables form the vertices. The four edges represent pairwise conditional associations. Edges do not connect X and Y or connect X and Z , the conditionally independent pairs.

[Figure 10.1](#) Conditional independence graphs for loglinear models (WX, WY, WZ, YZ) and (WX, WYZ) .



Two loglinear models with the same pairwise associations have the same conditional independence graph. For instance, the [Figure 10.1](#) graph is also the one for model (WX, WYZ) , which adds a three-factor WYZ interaction.

A set of properties, called *Markov properties*, allows us to deduce from the graph the conditional independence structure between variables and groups of variables. One such property, the *global Markov property*, links statements of conditional independence between two variables or groups of variables to the concept of separation in the graph. A *path* in a conditional independence graph is a sequence of edges leading from one variable to another. Two variables (or groups of variables) X and Y are said to be *separated* by a subset of variables if all paths connecting X and Y intersect that subset. For instance, in [Figure 10.1](#), W separates X and Y , since any path connecting X and Y goes through W . The subset $\{W, Z\}$ also separates X and Y . The global Markov property states that two variables are conditionally independent given *any* subset of variables that separates them (Darroch et al. 1980, Kreiner 1987). Thus, not only are X and Y conditionally independent given W and Z , but also given W alone. Similarly, X and Z are conditionally independent given W alone. This property is equivalent to a *local Markov property* according to which a variable is conditionally independent of all other variables, given its adjacent neighbors to which it's connected with an edge.

10.1.2 Graphical Loglinear Models

Darroch et al. (1980) used graph theory to represent hierarchical loglinear models having a conditional independence structure. Those models, called *graphical models*, are represented by undirected graphs such as just shown. In the graph, a maximally connected subset is called a *clique*. In [Figure 10.1](#), for example, the three variables W , Y , and Z form a clique, but any two of those three variables would not form one. The second clique of this graph is XW . For a graphical model, the generating classes that provide the sufficient statistics are the cliques. So, in [Figure 10.1](#), the joint distribution for W , Y , and Z is a sufficient statistic, and the graphical model corresponding to that graph is (XW, WYZ) rather than (XW, WY, WZ, YZ) .

The family of graphical models contains the family of decomposable models (Andersen 1974, Thm. 5). Recall that for decomposable models, the joint distribution factors into a product of marginal distributions and conditional distributions, and direct ML estimates exist. Not all graphical models are decomposable, however. An example is the loglinear model (WX, XY, YZ, ZW) , which has the graphical appearance of a square. This model exhibits two conditional independences (between W and Y and between X and Z) but does not have direct ML estimates. Tables 1 and 2 in Darroch et al. (1980) portray all the decomposable and nondecomposable graphical models of dimension ≤ 5 .

10.1.3 Collapsibility in Three-Way Contingency Tables

We have seen that conditional associations in partial tables usually differ from marginal associations. Under the *collapsibility conditions* given in Section 2.3.6, however, they are the same. For the odds ratio, the collapsibility conditions relate to logistic models, as we observed in Section 6.4.8, and loglinear models. Recall that in a three-way table, XY marginal and conditional odds ratios are identical if either Z and X are conditionally independent or if Z and Y are conditionally independent (or both). These conditions occur for loglinear models (XY, YZ) , (XY, XZ) , and (XY, Z) .

The proof of the collapsibility conditions follows directly from the model formulas. We illustrate with model (XY, XZ) . For it, the XY marginal table has

$$\begin{aligned}\mu_{ij+} &= \sum_k \exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}) \\ &= \exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}) \sum_k \exp(\lambda_k^Z + \lambda_{ik}^{XZ}).\end{aligned}$$

The loglinear model for that marginal table satisfies

$$\log \mu_{ij+} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} + \xi_i,$$

where $\xi_i = \log(\sum_k \exp(\lambda_k^Z + \lambda_{ik}^{XZ}))$ and can be combined with the λ_i^X term to give the main effect for X . Note that $\{\lambda_{ij}^{XY}\}$ are the same for the marginal table. Since the XY odds ratios are functions of $\{\lambda_{ij}^{XY}\}$, they are the same in both marginal and partial tables.

We illustrate for the student survey data ([Table 9.3](#)) from Section 9.2.4, about alcohol, cigarette, and marijuana use. Model (AM, CM) specifies AC conditional independence, given M . It has conditional independence graph

$$A \text{ --- } M \text{ --- } C.$$

Consider the AM association. Since C is conditionally independent of A , the AM fitted conditional odds ratios are the same as the AM fitted marginal odds ratio collapsed over C . From [Table 9.5](#), both equal 61.9. Similarly, the CM association is collapsible. The AC association is not, because M is conditionally dependent with both A and C in model (AM, CM) ; that is, an edge connects M to both A and C . Thus, A and C may be marginally dependent, even though they are conditionally independent. In fact, from [Table 9.5](#), the fitted AC marginal odds ratio for this model is 2.7 rather than 1.0.

For model (AC, AM, CM) , no pair is conditionally independent. No collapsibility conditions are fulfilled. [Table 9.5](#) showed that each pair has quite different fitted marginal and conditional associations for this model. When a model contains all two-factor effects, effects may change after collapsing over any variable.

10.1.4 Collapsibility for Multiway Tables

Bishop et al. (1975, p. 47) provided a parametric collapsibility condition for multiway tables:

Suppose that a model for a multiway table partitions variables into three mutually exclusive subsets, A , B , C , such that B separates A and C . After collapsing the table over the variables in C , parameters relating variables in A and parameters relating variables in A to variables in B are unchanged.

We illustrate using model (WX, WY, WZ, YZ) ([Figure 10.1](#)). Let $A = \{X\}$, $B = \{W\}$, and $C = \{Y, Z\}$. Since the XY and XZ terms do not appear, all parameters linking set A with set C equal zero, and B separates A and C . If we collapse over Y and Z , the WX association is unchanged. Next, identify $A = \{Y, Z\}$, $B = \{W\}$, $C = \{X\}$. Then, conditional associations among W , Y , and Z remain the same after collapsing over X .

This result also implies that when any variable is independent of all other variables, collapsing over it does not affect any other model terms. In model (WX, WY, XY, Z) , for instance, associations among W , X , and Y are the same as in model (WX, WY, XY) .

When the separating set B contains more than one variable, although parameter values are unchanged in collapsing over set C , the ML estimates of those parameters may differ slightly. A stronger collapsibility definition also requires that the estimates be identical. This condition of commutativity of fitting and collapsing holds if the model contains the highest-order term relating variables in B to each other; that is, it is *graphical*. Asmussen and Edwards (1983) discussed this property, which relates to decomposability.

10.2 MODEL SELECTION AND COMPARISON

Strategies for selecting and comparing loglinear models are similar to those for logistic regression presented in Section 6.1. A model should be complex enough to fit well but also relatively simple to interpret, smoothing rather than overfitting the data.

10.2.1 Considerations in Model Selection

The potentially useful models are usually a small subset of the possible models. A study designed to answer certain questions through confirmatory analyses may plan to compare models that differ only by the inclusion of certain terms. Also, models should recognize distinctions between response and explanatory variables. The modeling process should concentrate on terms linking responses and terms linking explanatory variables to responses. The model should contain the most general interaction term relating the explanatory variables. From the likelihood equations, this has the effect of equating the fitted totals to the sample totals at combinations of their levels. This is natural, since we normally treat such totals as fixed. Related to this, certain marginal totals are often fixed by the sampling design. Any potential model should include those totals as sufficient statistics, so likelihood equations equate them to the fitted totals.

Consider [Table 9.8](#) with I = automobile injury and S = seat-belt use as response variables. Since G = gender and L = location are explanatory variables, we treat $\{n_{g+\ell+}\}$ as fixed at each combination for G and L . For example, 20,629 women had accidents in urban locations, so the fitted counts should have 20,629 women in urban locations. To ensure this, a loglinear model should contain the GL term, which implies from its likelihood equations that $\{\mu_{g+\ell+} = n_{g+\ell+}\}$. Thus, the model should be at least as complex as (GL, S, I) and focus on the effects of G and L on S and I as well as the SI association.

For exploratory studies, one approach first fits the model having single-factor terms, then the model having two-factor and single-factor terms, then the model having three-factor and lower terms, and so on. Fitting such models often reveals a restricted range of good-fitting models. In Section 9.4.2 we used this strategy with the automobile injury data set. Automatic search mechanisms among possible models, such as backward elimination, may also be useful. However, they should be used with care and skepticism, as they need not yield a meaningful model.

10.2.2 Example: Model Building for Student Survey

The study on the use of alcohol (A), cigarettes (C), and marijuana (M) by a sample of high school seniors also classified students by gender (G) and race (R). [Table 10.1](#) shows the five-dimensional contingency table. In selecting a model, we treat A , C , and M as response variables and G and R as explanatory. Thus, a model should contain the GR term, which forces the GR fitted marginal totals to equal the sample marginal totals.

[Table 10.1](#) Alcohol, Cigarette, and Marijuana Use, by Gender and Race

Alcohol Use	Cigarette Use	Marijuana Use							
		Race = White				Race = Other			
		Female		Male		Female		Male	
Yes	Yes	405	268	453	228	23	23	30	19
No	Yes	13	218	28	201	2	19	1	18
	No	1	17	1	17	0	1	1	8
		1	117	1	133	0	12	0	17

Source: Harry Khamis, Wright State University.

[Table 10.2](#) displays goodness-of-fit tests for several models. Because many cell counts are small, the chi-squared approximation for G^2 may be poor, but this index is useful for comparing models. The first model listed contains only the GR association and assumes conditional independence for the other nine pairs of associations. It fits horribly, which is no surprise. Model 2, with all two-factor terms, seems to fit well. Model 3, containing all the three-factor interaction terms, also fits well, but the improvement in fit is not great (difference in G^2 of $15.3 - 5.3 = 10.0$ based on $df = 16 - 6 = 10$). Thus, we consider models without three-factor terms. Beginning with model 2, we eliminate two-factor terms. We use backward elimination, sequentially taking out terms for which the resulting increase in G^2 is smallest, when refitting the model.

[Table 10.2](#) Goodness-of-Fit Tests for Loglinear Models for [Table 10.1](#)

Model ^a	G^2	χ^2	df
1. Mutual independence + GR	1325.14	1454.14	25
2. Homogeneous association	15.34	18.68	16
3. All three-factor terms	5.27	4.80	6
4a. (2)– AC	201.20	190.60	17
4b. (2)– AM	106.96	108.11	17
4c. (2)– CM	513.47	474.26	17
4d. (2)– AG	18.72	23.14	17
4e. (2)– AR	20.32	30.32	17
4f. (2)– CG	16.32	19.16	17
4g. (2)– CR	15.78	20.12	17
4h. (2)– GM	25.16	27.97	17
4i. (2)– MR	18.93	22.83	17
5. (AC, AM, CM, AG, AR, GM, GR, MR)	16.74	20.51	18
6. (AC, AM, CM, AG, AR, GM, GR)	19.91	23.02	19
7. (AC, AM, CM, AG, AR, GR)	28.81	32.13	20

^a G , gender; R , race; A , alcohol use; C , cigarette use; M , marijuana use.

[Table 10.2](#) shows the start of this process. Nine pairwise associations are candidates for removal from model 2 (all except GR), shown in models 4a through 4i. The smallest increase in G^2 , compared with model 2, occurs in removing the CR term (i.e., model 4g). The increase is $15.78 - 15.34 = 0.44$, with $df = 17 - 16 = 1$, so this elimination seems sensible. After removing it, the smallest additional increase results from removing the CG term (model 5), resulting in $G^2 = 16.74$ with $df = 18$. Removing next the MR term (model 6) yields $G^2 = 19.91$ with $df = 19$.

Further removals have a more severe effect. For instance, removing the AG term increases G^2 by 5.26, with $df = 1$. Ordinary P -values do not apply for such statistics, since the data suggested these tests, but it seems safest not to drop additional terms. [See Westfall and Wolfinger (1997) and Westfall and Young (1993) for methods of adjusting P -values to account for multiple tests.] Model 6, denoted by (AC, AM, CM, AG, AR, GM, GR), has conditional independence graph



Every path between C and $\{G, R\}$ involves a variable in $\{A, M\}$. Given the outcome on alcohol use and marijuana use, the model states that cigarette use is independent of both gender and race. Collapsing over the explanatory variables race and gender, the conditional associations between C and A and between C and M are the same as with the model (AC, AM, CM) fitted in Section 9.2.4.

Removing the GM term from this model yields model 7 in [Table 10.2](#). Its graph reveals that A separates $\{G, R\}$ from $\{C, M\}$. Thus, all pairwise conditional associations among A , C , and M in model 7 are identical to those in model (AC, AM, CM) , collapsing over G and R . In fact, model 7 does not fit all that badly ($G^2 = 28.81$ with $df = 20$) considering the large sample size. So, we could collapse over gender and race in studying associations among the primary variables. An advantage of the full five-variable model is that it estimates effects of gender and race on these responses, in particular, the effects of race and gender on alcohol use and the effect of gender on marijuana use.

10.2.3 Loglinear Model Comparison Statistics

Consider two loglinear models, M_1 and M_0 , with M_0 a special case of M_1 . In comparing pairs of models above, we used the likelihood-ratio statistic for testing M_0 against M_1 , $G^2(M_0|M_1) = G^2(M_0) - G^2(M_1)$.

Let \mathbf{n} denote a column vector of the observed cell counts $\{n_i\}$. Let $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ denote vectors of the fitted values $\{\hat{\mu}_{0i}\}$ and $\{\hat{\mu}_{1i}\}$ for M_0 and M_1 . The deviance $G^2(M_0)$ for the simpler model partitions into

$$(10.1) \quad G^2(M_0) = G^2(M_1) + G^2(M_0|M_1).$$

Just as $G^2(M)$ measures the distance of fitted values for M from \mathbf{n} , $G^2(M_0|M_1)$ measures the distance of fit $\boldsymbol{\mu}_0$ from fit $\boldsymbol{\mu}_1$. In this sense, decomposition (10.1) expresses a certain orthogonality: The distance of \mathbf{n} from $\boldsymbol{\mu}_0$ equals the distance of \mathbf{n} from $\boldsymbol{\mu}_1$ plus the distance of $\boldsymbol{\mu}_1$ from $\boldsymbol{\mu}_0$.

As noted in Section 4.5.4, the model comparison statistic simplifies to

$$(10.2) \quad G^2(M_0|M_1) = 2 \sum_i n_i \log(\hat{\mu}_{1i}/\hat{\mu}_{0i}).$$

The two loglinear models have the matrix form (9.19), namely,

$$\log \boldsymbol{\mu}_0 = \mathbf{X}_0 \boldsymbol{\beta}_0 \quad \text{and} \quad \log \boldsymbol{\mu}_1 = \mathbf{X}_1 \boldsymbol{\beta}_1.$$

Since M_0 is simpler than M_1 , we can express $\log \boldsymbol{\mu}_0 = \mathbf{X}_0 \boldsymbol{\beta}_0 = \mathbf{X}_1 \boldsymbol{\beta}_1^*$, where $\boldsymbol{\beta}_1^*$ equals $\boldsymbol{\beta}_0$ with 0 elements appended corresponding to the extra parameters in $\boldsymbol{\beta}_1$ that are not in $\boldsymbol{\beta}_0$. Then, from (10.2),

$$(10.3) \quad \begin{aligned} G^2(M_0|M_1) &= 2\mathbf{n}^T (\log \boldsymbol{\mu}_1 - \log \boldsymbol{\mu}_0) = 2\mathbf{n}^T [\mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 - \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1^*] \\ &= 2\hat{\boldsymbol{\mu}}_1^T [\mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 - \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1^*] = 2\hat{\boldsymbol{\mu}}_1^T (\log \boldsymbol{\mu}_1 - \log \boldsymbol{\mu}_0) \\ &= 2 \sum_i \hat{\mu}_{1i} \log(\hat{\mu}_{1i}/\hat{\mu}_{0i}), \end{aligned}$$

where the replacement of \mathbf{n} by $\boldsymbol{\mu}_1$ follows from the likelihood equations $\mathbf{n}^T \mathbf{X}_1 = \hat{\boldsymbol{\mu}}_1^T \mathbf{X}_1$ for M_1 [recall (9.21)]. Statistic (10.3) has the same form as $G^2(M_0)$, but with $\{\hat{\mu}_{1i}\}$ playing the role of the observed data. Simon (1973) showed a general result of this type for natural exponential family distributions. From Section 4.5.5, the Pearson statistic $X^2(M_0|M_1)$ for comparing loglinear models is $\sum_i (\hat{\mu}_{1i} - \hat{\mu}_{0i})^2 / \hat{\mu}_{0i}$, which has the usual Pearson form with $\{\hat{\mu}_{1i}\}$ in place of $\{n_i\}$.

When M_0 holds, $G^2(M_0)$ and $G^2(M_1)$ have large-sample chi-squared distributions, and $G^2(M_0|M_1)$ is asymptotically chi-squared with df equal to the difference between df for M_0 and M_1 . Haberman (1977a) showed that $G^2(M_0|M_1)$ and $X^2(M_0|M_1)$ have the same null large-sample behavior, even for fairly sparse tables. (Under certain conditions, their difference converges in probability to 0 as n increases.) When M_1 holds but M_0 does not, $G^2(M_1)$ still has its asymptotic chi-squared distribution, but the other two statistics tend to grow unboundedly as n increases.

10.2.4 Partitioning Chi-Squared with Model Comparisons

[Equation \(10.1\)](#) utilizes the property by which a chi-squared statistic with $df > 1$ partitions into components. We used such partitionings in tests for trend with ordinal variables, such as in linear logit and linear probability models (Section 5.3.5). More generally, this property applies with a set of nested models to test a sequence of hypotheses. The separate tests for comparing pairs of models are asymptotically independent.

For example, a chi-squared decomposition with $J - 1$ models justifies the partitioning of G^2 stated in Section 3.3.3 for testing independence in $2 \times J$ tables. For $j = 2, \dots, J$, let M_j denote the model that satisfies

$$\theta_i = (\mu_{1i} \mu_{2,i+1}) / (\mu_{1,i+1} \mu_{2i}) = 1, \quad i = 1, \dots, j - 1.$$

For M_j , the $2 \times j$ table consisting of columns 1 through j satisfies independence. Model M_J is independence in the complete $2 \times J$ table. Model M_h is a special case of M_j whenever $h > j$. By [\(10.2\)](#),

$$\begin{aligned} G^2(M_J) &= G^2(M_J | M_{J-1}) + G^2(M_{J-1}) \\ &= G^2(M_J | M_{J-1}) + G^2(M_{J-1} | M_{J-2}) + G^2(M_{J-2}) \\ &= \dots = G^2(M_J | M_{J-1}) + \dots + G^2(M_3 | M_2) + G^2(M_2). \end{aligned}$$

From [\(10.3\)](#), $G^2(M_j | M_{j-1})$ has the G^2 form with the fitted values for model M_{j-1} playing the role of the observed data. Substitution of fitted values for the two models into [\(10.3\)](#) shows that $G^2(M_j | M_{j-1})$ is identical to G^2 for testing independence in a 2×2 table; the first column combines column 1 through $j - 1$ of the original table, and the second column is column j of the original table.

With several preplanned comparisons, simultaneous test procedures lessen the probability of attributing importance to sample effects that merely reflect chance variation. These procedures use adjusted significance levels. For a set of s independent tests for nested models, when each test has approximate size $1 - (1 - \alpha)^{1/s}$, the overall asymptotic $P(\text{type I error}) \asymp \alpha$. For instance, suppose that we test the fit of (WXZ, WY, XY, ZY) , compare that model to (WX, WZ, XZ, WY, XY, ZY) , and compare that model to (WX, WZ, XZ, WY, ZY) . To have overall $\alpha = 0.05$ for the $s = 3$ tests, use level $1 - (0.95)^{1/3} = 0.01695$ for each.

10.2.5 Identical Marginal and Conditional Tests of Independence

A test using $G^2(M_0|M_1)$ simplifies dramatically when both models have direct estimates. In that case, the models have independence linkages necessary to ensure collapsibility. A test of conditional independence then has the same result as the test of independence applied to the marginal table. Sundberg (1975) proved the following: When two direct models M_0 and M_1 are identical except for a pairwise association term, $G^2(M_0|M_1)$ is identical to G^2 for testing independence in the marginal table for that pair of variables. Bishop (1971) and Goodman (1970, 1971b) have related discussion.

For instance, $G^2[(X, Y, Z)|(XY, Z)]$ tests $\lambda^{XY} = 0$ in model (XY, Z) . Thus, it tests XY conditional independence under the assumption that X and Y are jointly independent of Z . Using the two sets of fitted values, from (10.3), it equals

$$\begin{aligned} & 2 \sum_i \sum_j \sum_k \frac{n_{ij} + n_{++k}}{n} \log \frac{n_{ij} + n_{++k}/n}{n_{i++} n_{+j+} n_{++k}/n^2} \\ &= 2 \sum_i \sum_j n_{ij+} \log \frac{n_{ij+}}{n_{i++} n_{+j+}/n}, \end{aligned}$$

which is $G^2[(X, Y)]$ for testing independence in the marginal XY table. This is not surprising. The collapsibility conditions imply that for model (XY, Z) , the marginal XY association is the same as the conditional XY association.

10.3 RESIDUALS FOR DETECTING CELL-SPECIFIC LACK OF FIT

The model comparison test using $G^2(M_0|M_1)$ is useful for detecting whether an extra term improves a model fit. Cell residuals provide a cell-specific indication of model lack of fit.

10.3.1 Residuals for Loglinear Models

In Section 4.5.6 we presented residuals that apply to any Poisson GLM. For cell i in a contingency table with observed count n_i and fitted value $\hat{\mu}_i$, the *Pearson residual* is

$$(10.4) \quad e_i = \frac{n_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}.$$

These relate to the Pearson statistic by $\sum_i e_i^2 = X^2$.

The corresponding standardized residual (Haberman 1973a) is

$$r_i = e_i / \sqrt{1 - h_i},$$

where the leverage h_i is a diagonal element of the estimated hat matrix. This has an asymptotic standard normal distribution and is preferable to the Pearson residual. Alternative residuals use components of the deviance.

10.3.2 Example: Student Survey Revisited

For [Table 10.1](#) cross-classifying alcohol, cigarette, and marijuana use by gender and race, we suggested in Section 10.2.2 that the model with all two-factor associations is plausible. For it, the only large standardized residual equals 3.2, resulting from a fitted value of 3.1 in the cell having a count of 8. Further comparisons suggested that the simpler model ($AC, AM, CM, AG, AR, GM, GR$) is adequate. Its only large standardized residual equals 3.3, from the fitted value of 2.9 in that cell.

From these standardized residuals, the number of nonwhite males who did not use alcohol or marijuana but who smoked cigarettes is somewhat greater than either model predicts. The residuals do not suggest problems with either model, considering the large sample size, and the many cells studied.

10.3.3 Identical Loglinear and Logistic Standardized Residuals

In Section 9.5 we showed that logistic models for contingency tables are equivalent to certain loglinear models. However, a Pearson residual for a logistic model differs from a Pearson residual for a loglinear model. The numerators comparing the i th observed and fitted binomial or Poisson count are the same, since the model fitted values are the same. However, the logistic model uses a fitted binomial standard deviation in the denominator [see [\(6.1\)](#)], whereas the loglinear model uses a fitted Poisson standard deviation [see [\(10.4\)](#)]. Thus, the logistic Pearson residual exceeds the loglinear Pearson residual.

Once divided by estimated standard errors, the standardized residuals are identical for the two models. This is another reason for preferring standardized residuals over ordinary Pearson residuals.

10.4 MODELING ORDINAL ASSOCIATIONS

The loglinear models presented so far have a serious limitation—they treat all classifications as nominal. If the order of a variable's categories changes in any way, the fit is the same. For ordinal classifications, these models ignore important information.

Refer to [Table 10.3](#). Subjects were asked their opinion about a man and woman having sexual relations before marriage and also asked whether methods of birth control should be available to teenagers between the ages of 14 and 16. For the loglinear model of independence, denoted by I , $G^2(I) = 127.65$ with $df = 9$. The model fits poorly. Yet, adding the ordinary association term makes it saturated and unhelpful.

Table 10.3 Opinions About Premarital Sex and Availability of Teenage Birth Control

Premarital Sex	Teenage Birth Control			
	Strongly Disagree	Disagree	Agree	Strongly Agree
Always wrong	81 (42.4) ^a (80.9) ^b 7.6 ^c	68 (51.2) (67.6) 3.1	60 (86.4) (69.4) -4.1	38 (67.0) (29.1) -4.8
Almost always wrong	24 (16.0) (20.8) 2.3	26 (19.3) (23.1) 1.8	29 (32.5) (31.5) -0.8	14 (25.2) (17.6) -2.8
Wrong only sometimes	18 (30.1) (24.4) -2.7	41 (36.3) (36.1) 1.0	74 (61.2) (65.7) 2.2	42 (47.4) (48.8) -1.0
Not wrong at all	36 (70.6) (33.0) -6.1	57 (85.2) (65.1) -4.6	161 (143.8) (157.4) 2.4	157 (111.4) (155.5) 6.8

^aIndependence model fit.

^bLinear-by-linear association model fit.

^cStandardized residuals for the independence model fit.

Source: 1991 General Social Survey, National Opinion Research Center.

[Table 10.3](#) also contains fitted values and standardized residuals for independence. The residuals in the corners stand out. Sample counts are much larger than independence predicts where both responses are the most negative possible or the most positive possible. By contrast, the counts are much smaller than fitted values where one response is the most positive and the other is the most negative. Cross-classifications of ordinal variables often exhibit their greatest deviations from independence in the corner cells. This pattern for [Table 10.3](#) indicates lack of fit in the form of a positive trend. People who are more willing to make birth control available to teenagers also tend to feel more tolerant about premarital sex.

Models for ordinal variables use association terms that permit trends. The models are more complex than the independence model, yet unsaturated. They are called *association models*, because they focus on the association structure. Tests with association models also have improved power for detecting trends.

10.4.1 Linear-by-Linear Association Model for Two-Way Tables

For two-way contingency tables, a simple model for two ordinal variables assigns ordered row scores $u_1 \leq u_2 \leq \dots \leq u_I$ and column scores $v_1 \leq v_2 \leq \dots \leq v_J$. The model is

$$(10.5) \log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j,$$

with constraints such as $\lambda_I^X = \lambda_J^Y = 0$. This is the special case of the saturated model (9.2) in which $\lambda_{ii}^{XY} = \beta u_i v_j$. It uses only one parameter to describe association, whereas the saturated model uses $(I-1)(J-1)$ parameters.

Independence occurs when $\beta = 0$. The term $\beta u_i v_j$ represents the deviation of $\log \mu_{ij}$ from independence. The deviation is linear in the Y scores at a fixed level of X and linear in the X scores at a fixed level of Y . In column j , for instance, the deviation is a linear function of X , having form (slope) \times (score for X), with slope βv_j . Because of this property, (10.5) is called the *linear-by-linear association model* (abbreviated, $L \times L$). The model has its greatest departures from independence in the corners of the table.

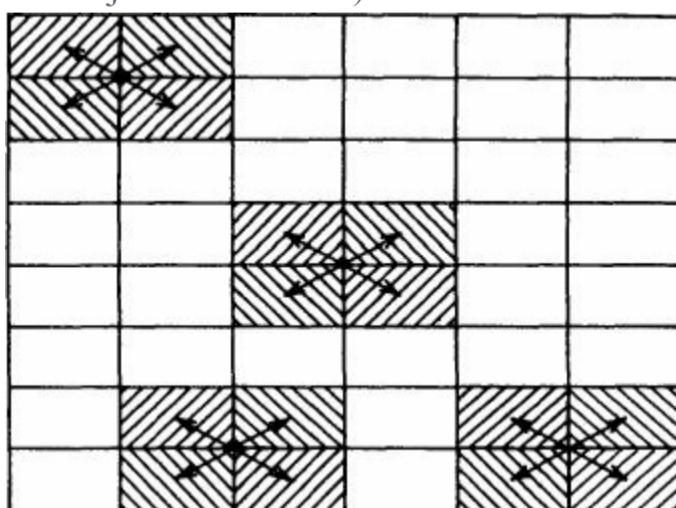
The direction and strength of the association depend on β . When $\beta > 0$, Y tends to increase as X increases. Expected frequencies are larger than expected (under independence) in cells where X and Y are both high or both low. When $\beta < 0$, Y tends to decrease as X increases. When the data display a positive or negative trend, the $L \times L$ model usually fits much better than the independence model.

For the 2×2 table using the cells intersecting rows a and c with columns b and d , direct substitution shows that the model has

$$(10.6) \log \frac{\mu_{ab}\mu_{cd}}{\mu_{ad}\mu_{cb}} = \beta(u_c - u_a)(v_d - v_b).$$

This log odds ratio is stronger as $|\beta|$ increases and for pairs of categories that are farther apart. Simple interpretations result when $u_2 - u_1 = \dots = u_I - u_{I-1}$ and $v_2 - v_1 = \dots = v_J - v_{J-1}$. When $\{u_i = i\}$ and $\{v_j = j\}$, for instance, the *local odds ratios* (2.10) for adjacent rows and adjacent columns have common value e^β . Goodman (1979a) called this case *uniform association*. Figure 10.2 portrays local odds ratios having uniform value.

Figure 10.2 Constant odds ratio implied by uniform association model. (Note: β = the constant log odds ratio for adjacent rows and adjacent columns.)



The choice of scores affects the interpretation of β . Often, the response scale discretizes an inherently continuous scale. It is sensible to choose scores that approximate distances between midpoints of categories for the underlying scale. It is sometimes useful to standardize the scores, subtracting the mean and dividing by the standard deviation, so

$$\sum_i u_i \pi_{i+} = \sum_j v_j \pi_{+j} = 0,$$

$$\sum_i u_i^2 \pi_{i+} = \sum_j v_j^2 \pi_{+j} = 1.$$

The $L \times L$ model tends to fit well when an underlying continuous distribution is approximately

bivariate normal. For standardized scores, β is then comparable to $\rho/(1 - \rho^2)$, where ρ is the underlying correlation (Goodman 1981a,b, 1985). For weak associations, $\beta \approx \rho$.

10.4.2 Corresponding Logistic Model for Adjacent Responses

A logistic formulation of the $L \times L$ model treats Y as a response and X as explanatory. Let $\pi_{j|i} = P(Y = j|X = i)$. Using logits for adjacent response categories (Section 8.3.4),

$$\log \frac{\pi_{j+1|i}}{\pi_{j|i}} = \log \frac{\mu_{i,j+1}}{\mu_{ij}} = (\lambda_{j+1}^r - \lambda_j^r) + \beta(v_{j+1} - v_j)u_i.$$

For unit-spaced $\{v_j\}$, this simplifies to

$$\log \frac{\pi_{j+1|i}}{\pi_{j|i}} = \alpha_j + \beta u_i,$$

where $\alpha_j = \lambda_{j+1}^r - \lambda_j^r$. The same linear logit effect β applies simultaneously for all $(J - 1)$ pairs of adjacent response categories: The odds that $Y = j + 1$ instead of $Y = j$ multiply by e^β for each unit change in X . In using equal-interval response scores, we implicitly assume that the effect of X is the same on each of the $J - 1$ adjacent-categories logits for Y .

10.4.3 Likelihood Equations and Model Fitting

The Poisson log likelihood $L(\boldsymbol{\mu}) = \sum_i \sum_j n_{ij} \log \mu_{ij} - \sum_i \sum_j \mu_{ij}$ for a two-way table simplifies for the $L \times L$ model (10.5) to

$$L(\boldsymbol{\mu}) = n\lambda + \sum_i n_{i+} \lambda_i^X + \sum_j n_{+j} \lambda_j^Y + \beta \sum_i \sum_j u_i v_j n_{ij} \\ - \sum_i \sum_j \exp(\lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j).$$

Differentiating $L(\boldsymbol{\mu})$ with respect to $(\lambda_i^X, \lambda_j^Y, \beta)$ and setting the three partial derivatives equal to zero yields likelihood equations

$$\hat{\mu}_{i+} = n_{i+}, \quad i = 1, \dots, I, \quad \hat{\mu}_{+j} = n_{+j}, \quad j = 1, \dots, J,$$

$$\sum_i \sum_j u_i v_j \hat{\mu}_{ij} = \sum_i \sum_j u_i v_j n_{ij}.$$

Iterative methods such as Newton–Raphson yield the ML fit.

Let $p_{ij} = n_{ij}/n$ and $\hat{\pi}_{ij} = \hat{n}_{ij}/n$. The third likelihood equation implies that

$$\sum_i \sum_j u_i v_j \hat{\pi}_{ij} = \sum_i \sum_j u_i v_j p_{ij}.$$

Since marginal distributions and hence marginal means and variances are identical for fitted and observed distributions, the third equation implies that the correlation between the scores for X and Y is the same for both distributions. The fitted counts display the same positive or negative trend as the data.

Since $\{u_i\}$ and $\{v_j\}$ are fixed, the $L \times L$ model (10.5) has only one more parameter (β) than the independence model. Its residual df = $IJ - I - J$. It is unsaturated for all but 2×2 tables.

10.4.4 Example: Sex and Birth Control Opinions Revisited

[Table 10.3](#) also reports fitted values for the linear-by-linear association model, using scores (1, 2, 3, 4) for rows and columns. [Table 10.4](#) shows software output. To get this, we added to the independence model a variable (denoted by “linlin”) having values equal to the product of row and column numbers. Compared with the independence model, for which $G^2(I) = 127.65$ with $df = 9$, the $L \times L$ model fits dramatically better [$G^2(L \times L) = 11.53$, $df = 8$]. This is especially noticeable in the corners, where it predicts the greatest departures from independence.

Table 10.4 Linear-by-Linear Association Model Output (SAS) for [Table 10.3](#)

Criteria For Assessing Goodness Of Fit						
Criterion	DF	Value				
Deviance	8	11.5337				
Pearson Chi-Square	8	11.5085				
Parameter	Estimate	Standard Error	Wald 95% Conf. Limits	Chi-Square	Pr > ChiSq	
Intercept	0.4735	0.4339	-0.3769 1.3239	1.19	0.2751	
premar 1	1.7537	0.2343	1.2944 2.2129	56.01	<.0001	
premar 2	0.1077	0.1988	-0.2820 0.4974	0.29	0.5880	
premar 3	-0.0163	0.1264	-0.2641 0.2314	0.02	0.8972	
premar 4	0.0000	0.0000	0.0000 0.0000	.	.	
birth 1	1.8797	0.2491	1.3914 2.3679	56.94	<.0001	
birth 2	1.4156	0.1996	1.0243 1.8068	50.29	<.0001	
birth 3	1.1551	0.1291	0.9021 1.4082	80.07	<.0001	
birth 4	0.0000	0.0000	0.0000 0.0000	.	.	
linlin	0.2858	0.0282	0.2305 0.3412	102.46	<.0001	
LR Statistics						
Source	DF	Chi-Square	Pr > ChiSq			
linlin	1	116.12	>.0001			

The ML estimate $\hat{\beta} = 0.286$ ($SE = 0.028$) indicates that subjects having more favorable attitudes about teen birth control also tend to have more tolerant attitudes about premarital sex. The estimated local odds ratio is $\exp(\hat{\beta}) = \exp(0.286) = 1.33$. The 95% Wald confidence interval is $\exp[0.286 \pm 1.96(0.028)]$, or (1.26, 1.41). The strength of association seems weak. From [\(10.6\)](#), however, nonlocal odds ratios are stronger. The estimated odds ratio for the four corner cells, obtained using $\hat{\beta}$ or the corner fitted values, equals

$$\exp[\hat{\beta}(u_4 - u_1)(v_4 - v_1)] = \exp[0.286(4 - 1)(4 - 1)] = \frac{80.9 \times 155.5}{29.1 \times 33.0} = 13.1.$$

Suppose we regard categories 2 and 3 as farther apart than categories 1 and 2, or categories 3 and 4. Scores such as {1, 2, 4, 5} for rows and columns recognize this. The $L \times L$ model then has $G^2 = 8.85$ ($df = 8$) and $\hat{\beta} = 0.146$ ($SE = 0.014$). But we need not regard the scores as approximations for distances between categories or as reasonable scalings of ordinal variables in order for the models to be valid. They merely imply a certain pattern for the odds ratios. If the $L \times L$ model fits well with equally spaced row and column scores, the uniform local odds ratio describes the association regardless of whether the scores are sensible indexes of true distances between categories.

For scores $\{u_i = i\}$ with [Table 10.3](#), the marginal mean and standard deviation for premarital sex are 2.81 and 1.26. The standardized scores are $\{(i - 2.81)/1.26\}$, or (-1.44, -0.65, 0.15, 0.95). The standardized equal-interval scores for birth control are (-1.65, -0.69, 0.27, 1.23). For these scores, $\hat{\beta} = 0.374$. Solving $\hat{\beta} = \hat{\rho}/(1 - \hat{\rho}^2)$ for $\hat{\rho}$ yields $\hat{\rho} = 0.333$. If there is an underlying bivariate normal distribution, we estimate the correlation to be 0.333.

10.4.5 Directed Ordinal Test of Independence

For the linear-by-linear association model, H_0 : independence is $H_0: \beta = 0$. The likelihood-ratio test statistic equals

$$G^2(I|L \times L) = G^2(I) - G^2(L \times L).$$

Designed to detect positive or negative trends, it has $df = 1$. For [Table 10.3](#), $G^2(I|L \times L) = 127.65 - 11.53 = 116.1$ has $P < 0.0001$, extremely strong evidence of an association. The Wald statistic $z^2 = (\hat{\beta}/SE)^2 = (0.286/0.0282)^2 = 102.5$ ($df = 1$) also shows strong evidence. The correlation statistic [\(3.16\)](#) presented in Section 3.4.1 for testing independence is the score statistic for $H_0: \beta = 0$ in this model (Exercise 10.27). It equals 112.6 ($df = 1$).

When the $L \times L$ model holds, the ordinal test using $G^2(I|L \times L)$ is asymptotically more powerful than the test using $G^2(I)$. This is because the power of a chi-squared test increases when df decrease, for fixed noncentrality (Section 5.3.8). When the $L \times L$ model holds, the noncentrality is the same for $G^2(I|L \times L)$ and $G^2(I)$; thus $G^2(I|L \times L)$ is more powerful, since its $df = 1$ compared with $(I-1)(J-1)$ for $G^2(I)$. The power advantage increases as I and J increase, since the noncentrality remains focused on $df = 1$ for $G^2(I|L \times L)$ but df also increases for $G^2(I)$.

10.4.6 Row Effects and Column Effects Association Models

Generalizations of the linear-by-linear association model treat some or all scores as parameters rather than fixed. To illustrate, replacing the ordered row values $\{\beta u_i\}$ in the linear-by-linear term $\beta u_i v_j$ in model (10.5) by unordered parameters (μ_i) gives

$$(10.7) \quad \log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \mu_i v_j.$$

Constraints are needed such as $\lambda_I^X = \lambda_J^Y = \mu_I = 0$. The $\{\mu_i\}$ are called *row effects* and the model is called the *row effects model*. Since the row effects are unordered but $\{v_j\}$ are ordered, this model treats X as nominal and Y as ordinal. The independence model is the special case $\mu_1 = \dots = \mu_I$. A corresponding *column effects model* has association term $u_i v_j$. It treats X as ordinal with ordered scores $\{u_i\}$ and Y as nominal with unordered parameters $\{v_j\}$.

The likelihood equations for the row effects model (10.7) are $\{\hat{\mu}_{i+} = n_{i+}\}$, $\{\hat{\mu}_{+j} = n_{+j}\}$, and

$$\sum_j v_j \hat{\mu}_{ij} = \sum_j v_j n_{ij}, \quad i = 1, \dots, I.$$

Let $\hat{\pi}_{j|i} = \hat{\mu}_{ij}/\hat{\mu}_{i+}$ and $p_{j|i} = n_{ij}/n_{i+}$. Since $\hat{\mu}_{i+} = n_{i+}$, the third likelihood equation is $\sum_j v_j \hat{\pi}_{j|i} = \sum_j v_j p_{j|i}$. For the conditional distribution within each row, the mean column score is the same for the fitted and sample distributions. The likelihood equations are solved using iterative methods such as Newton–Raphson.

With $\{v_{j+1} - v_j = 1\}$, the row effects model has adjacent-categories logit form

$$(10.8) \quad \log \frac{P(Y = j+1|X = i)}{P(Y = j|X = i)} = \alpha_j + \mu_i.$$

The effect in row i is identical for each pair of adjacent responses. Differences among $\{\mu_i\}$ compare rows with respect to their conditional distributions on Y . When $\mu_i = \mu_h$, rows h and i have identical conditional distributions. If $\mu_i > \mu_h$, Y is stochastically higher in row i than row h (Exercise 10.25). The row effects model also applies when X is ordinal but row scores for a linear-by-linear association structure are unknown and need to be estimated, as illustrated in the following example.

10.4.7 Example: Estimating Category Scores for Premarital Sex

In modeling [Table 10.3](#), we assigned equally spaced scores to both ordinal variables. If we regard those scores as a scaling with relevant distances between categories, it is not clear that this is sensible for categories such as (always wrong, almost always wrong, wrong only sometimes, not wrong at all). We next fitted the row effects model, again using column scores (1, 2, 3, 4). The model fits well, with deviance $G^2 = 7.59$ on $df = 6$, a decrease of 3.95 on $df = 2$ compared with the $L \times L$ model.

With constraint $\mu_4 = 0$, the other three row effect estimates contrast the first three rows with those who responded “not wrong at all” on premarital sex. The ML estimates are $\mu_1 = -0.584$ ($SE = 0.059$), $\mu_2 = -0.496$ ($SE = 0.080$), $\mu_3 = -0.203$ ($SE = 0.065$). So, the conditional distributions on the column variable differ more between rows 2 and 3 than between rows 1 and 2 or rows 3 and 4. This explains why the $L \times L$ model fitted better using scores (1, 2, 4, 5) than (1, 2, 3, 4). The further μ_i falls in the negative direction, the greater the tendency for those in row i to locate at the “strongly disagree” end of the column scale, relative to those in row 4. From [\(10.8\)](#) the model predicts constant odds ratios for adjacent columns of teenage birth control. For instance, the estimated odds that those in row 4 responded in category $j + 1$ instead of j were $\exp(\mu_4 - \mu_1) = \exp(0.584) = 1.79$ times the corresponding estimated odds for those in row 1, $j = 1, 2, 3$.

10.4.8 Ordinal Variables in Models for Multiway Tables

Multidimensional tables with ordinal responses can use generalizations of association models. In three dimensions, the rich collection of models includes association models that are more parsimonious than the model (XY , XZ , YZ) which treats all variables as nominal-scale, and models permitting heterogeneous association that, unlike model (XYZ), are unsaturated.

Models for association that are special cases of (XY , XZ , YZ) replace association terms by structured terms that account for ordinality. For instance, when both X and Y are ordinal, alternatives to λ_{ij}^{XY} are a linear-by-linear term $\beta u_i v_j$, a row effects term $\mu_i v_j$, or a column effects term $u_i v_j$, these provide a stochastic ordering of conditional distributions within rows and within columns, or only within rows, or only within columns. With a linear-by-linear term,

$$(10.9) \quad \log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta u_i v_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

has conditional local odds ratios (9.13) that satisfy

$$\log \theta_{ij(k)} = \beta(u_{i+1} - u_i)(v_{j+1} - v_j) \quad \text{for all } k.$$

The association is the same in each partial table, with *homogeneous linear-by-linear XY association*.

When the association is heterogeneous, structured terms for ordinal variables make effects simpler to interpret than in the saturated model. For instance, the *heterogeneous linear-by-linear XY association model* adds a term $\beta_k u_i v_j$ to model (10.9), thereby allowing the XY association to change across levels of Z .

Tests of conditional independence of ordinal classifications can generalize $G^2(NL \times L)$. For instance, we can compare the XY conditional independence model (XZ , YZ) to the homogeneous linear-by-linear XY association model (10.9). It tests $\beta = 0$ in that model, with $df = 1$. This is an alternative to the ordinal test of conditional independence in Section 8.4.3. Like Mantel's score statistic (8.19), this statistic uses correlation information, since $\sum_k (\sum_i \sum_j u_i v_j n_{ijk})$ is the sufficient statistic for β in model (10.9). In fact, the Mantel statistic provides a score test of $H_0: \beta = 0$ in that model.

10.5 GENERALIZED LOGLINEAR AND ASSOCIATION MODELS, CORRELATION MODELS, AND CORRESPONDENCE ANALYSIS

We've just seen how to construct loglinear models that describe ordinal associations. Other generalizations, some of which do not have loglinear form, can describe associations for both ordinal and nominal variables.

10.5.1 Generalized Loglinear Model

In Section 9.6.2, we expressed loglinear models for cell counts \mathbf{n} and expected frequencies $\boldsymbol{\mu}$ as $\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. A *generalized loglinear model* that allows many additional models is

$$(10.10) \quad \mathbf{C} \log(\mathbf{A}\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$$

for matrices \mathbf{C} and \mathbf{A} . The ordinary loglinear model results when \mathbf{C} and \mathbf{A} are identity matrices. Other special cases include logistic models for binary or multicategory responses (Grizzle et al. 1969).

For instance, the loglinear model of independence for a 2×2 table is equivalent to a model by which the logit for Y is the same in each row of X (see Section 9.1.2). That logit model has form (10.10): \mathbf{A} is a 4×4 identity matrix, so $\mathbf{A}\boldsymbol{\mu}$ is the 4×1 vector $\boldsymbol{\mu} = (\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22})^T$; the product $\mathbf{C} \log(\mathbf{A}\boldsymbol{\mu})$ forms the logit in row 1 and the logit in row 2 using

$$\mathbf{C} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix};$$

then $\mathbf{X} = (1, 1)^T$ is a 2×1 matrix, and $\boldsymbol{\beta}$ is a single constant α , so $\mathbf{X}\boldsymbol{\beta}$ forms a common value for those two logits.

The generalized loglinear model (10.10) includes models for association that are not ordinary loglinear models. For example, for a $I \times J$ cross-classification of two ordinal variables, the odds ratio obtained by collapsing to a 2×2 table by combining rows 1 to i , rows $i + 1$ to I , columns 1 to j , and columns $j + 1$ to J , is called a *global odds ratio*. A *uniform association* model specifies a common value for the $(I - 1)(J - 1)$ population global odds ratios. This is an alternative to the uniform association model for local odds ratios that results from the linear-by-linear association model with equally spaced scores.

In Chapters 11 and 12 we use the generalized loglinear model for models outside the classes of GLMs studied thus far, such as models for marginal distributions of multivariate responses. Lang (1996a, 2004, 2005) has developed ML fitting methods and an R function for generalized loglinear models. His methods apply to an even broader family of models for contingency tables, called *multinomial Poisson homogeneous models*, that have the form $\mathbf{L}(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$ for a general link function \mathbf{L} .

10.5.2 Multiplicative Row and Column Effects Model

The linear-by-linear association ($L \times L$) model (10.5) is a special case of the row effects (R) model, which has parameter row scores, and the column effects (C) model, which has parameter column scores. These models are special cases of a more general association model with row *and* column parameter scores. Replacing $\{u_i\}$ and $\{v_j\}$ in the $L \times L$ model by parameters yields the *row and column effects (RC)* model (Goodman 1979a)

$$(10.11) \log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \beta \mu_i v_j.$$

This model is *not* loglinear, because the predictor is a multiplicative (rather than linear) function of parameters μ_i and v_j . Identifiability requires location and scale constraints on $\{\mu_i\}$ and $\{v_j\}$. The model treats classifications as nominal; the same fit results from any permutation of rows or of columns. Parameter interpretation is simplest when at least one variable is ordinal, through the local log odds ratios

$$\log \theta_{ij} = \beta(\mu_{i+1} - \mu_i)(v_{j+1} - v_j).$$

Although it may seem appealing to use parameters instead of arbitrary scores, the *RC* model presents complications that do not occur with loglinear models. The likelihood may not be concave and may have local maxima. Independence is a special case, but it is awkward to test independence using the *RC* model. Haberman (1981) showed that the null distribution of $G^2(I) — G^2(RC)$ is not chi-squared but rather that of the maximum eigenvalue from a Wishart matrix. Haberman (1995) provided fitting methods for association models including nonlinear models such as this one. Software is available¹.

Goodman (1985) expressed the association term in the saturated model in a form that generalizes the $\beta \mu_i v_j$ term in the *RC* model, namely,

$$(10.12) \lambda_{ij}^{XY} = \sum_{k=1}^M \beta_k \mu_{ik} v_{jk},$$

where $M = \min(I - 1, J - 1)$. The parameters satisfy constraints such as

$$\sum_i \mu_{ik} \pi_{i+} = \sum_j v_{jk} \pi_{+j} = 0 \quad \text{for all } k,$$

$$\sum_i \mu_{ik}^2 \pi_{i+} = \sum_j v_{jk}^2 \pi_{+j} = 1 \quad \text{for all } k,$$

$$(10.13) \sum_i \mu_{ik} \mu_{ih} \pi_{i+} = \sum_j v_{jk} v_{jh} \pi_{+j} = 0 \quad \text{for all } k \neq h.$$

When $\beta_k = 0$ for $k > M^*$, model (10.12) is called the *RC(M^*)* model. The *RC* model (10.11) is the case $M^* = 1$.

10.5.3 Example: Mental Health and Parents' SES

[Table 10.5](#) describes the relationship between child's mental impairment and parents' socioeconomic status for a sample of residents of Manhattan (Goodman 1979a). The *RC* model fits well ($G^2 = 3.57$, $df = 8$). For scaling [\(10.13\)](#), the ML estimates are $(-1.11, -1.12, -0.37, 0.03, 1.01, 1.82)$ for the row scores, $(-1.68, -0.14, 0.14, 1.41)$ for the column scores, and $\hat{\beta} = 0.17$. Nearly all estimated local log odds ratios are positive, indicating a tendency for mental health to be better at higher levels of parents' SES.

Table 10.5 Cross-Classification of Mental Health Status and Socioeconomic Status

Parents' Socioeconomic Status	Mental Health Status			Impaired
	Well	Mild	Moderate	
		Symptom Formation	Symptom Formation	
A (high)	64	94	58	46
B	57	94	54	40
C	57	105	65	60
D	72	141	77	94
E	36	97	54	78
F (low)	21	71	54	71

Source: Reprinted with permission from L. Srole et al. *Mental Health in the Metropolis: The Midtown Manhattan Study*. New York: NYU Press, 1978, p. 289.

Ordinal loglinear models also fit well. For equal-interval scores, $G^2(L \times L) = 9.89$ ($df = 14$). The statistic $G^2(L \times L|RC) = 6.32$ ($df = 6$) tests that row and column scores in the *RC* model are equal-interval. The parameter scores do not provide a significantly better fit. It is sufficient to use a uniform local odds ratio to describe the table. For unit-spaced scores, $\hat{\beta} = 0.091$ ($SE = 0.015$), so the fitted local odds ratio is $\exp(0.091) = 1.09$. There is strong evidence of positive association, but the degree of association is rather weak, at least locally.

10.5.4 Correlation Models

A *correlation model* for two-way tables has many features in common with the *RC model* (Goodman 1985, 1986). In a one-dimensional version, it is

$$(10.14) \quad \pi_{ij} = \pi_{i+} \pi_{+j} (1 + \lambda \mu_i v_j),$$

where $\{\mu_i\}$ and $\{v_i\}$ are score parameters satisfying

$$\sum_i \mu_i \pi_{i+} = \sum_j v_j \pi_{+j} = 0 \quad \text{and} \quad \sum_i \mu_i^2 \pi_{i+} = \sum_j v_j^2 \pi_{+j} = 1.$$

The parameter λ is the correlation between the scores for joint distribution (10.14).

The correlation model is also called the *canonical correlation model*, because ML estimates of the scores maximize the correlation for (10.14). The general canonical correlation model is, for $M = \min(I - 1, J - 1)$,

$$\pi_{ij} = \pi_{i+} \pi_{+j} \left(1 + \sum_{k=1}^M \lambda_k \mu_{ik} v_{jk} \right),$$

where $0 \leq \lambda_M \leq \dots \leq \lambda_1 \leq 1$ and with constraints such as in (10.13). The parameter λ_k is the correlation between $\{\mu_{ik}, i = 1, \dots, I\}$ and $\{v_{jk}, j = 1, \dots, J\}$. The $\{\mu_{i1}\}$ and $\{v_{j1}\}$ are standardized scores that maximize the correlation λ_1 for the joint distribution; $\{\mu_{i2}\}$ and $\{v_{j2}\}$ are standardized scores that maximize the correlation $\{\lambda_2\}$, subject to $\{\mu_{i1}\}$ and $\{\mu_{i2}\}$ being uncorrected and $\{v_{j1}\}$ and $\{v_{j2}\}$ being uncorrected, and so on.

Unsaturated models result from taking $\lambda_k = 0$ for $k > M^*$ with $M^* < M$. Gilula and Haberman (1986) and Goodman (1985) discussed ML fitting. When λ is close to zero in (10.14), Goodman (1981a, 1985, 1986) noted that ML estimates of λ and the score parameters are similar to those of β and the score parameters in the *RC model*. Correlation models can also use fixed scores instead of parameter scores.

Goodman discussed advantages of association models over correlation models. The correlation model is not defined for all possible combinations of score values because of the constraint $0 \leq \pi_{ij} \leq 1$, ML fitted values do not have the same marginal totals as the observed data, and the model is not simply generalizable to multiway tables.

10.5.5 Correspondence Analysis

Correspondence analysis is a graphical way to represent associations in two-way contingency tables. The rows and columns are represented by points on a graph, the positions of which indicate associations. Goodman (1985, 1986) noted that coordinates of the points are reparameterizations of $\{\mu_{ik}\}$ and $\{v_{jk}\}$ in the general canonical correlation model. Correspondence analysis uses adjusted scores

$$x_{ik} = \lambda_k \mu_{ik}, \quad y_{jk} = \lambda_k v_{jk}.$$

These are close to zero for dimensions k in which the correlation λ_k is close to zero. A correspondence analysis graph uses the first two dimensions, plotting (x_{i1}, x_{i2}) for each row and (y_{j1}, y_{j2}) for each column.

Goodman (1985, 1986) used [Table 10.5](#) to illustrate the similarities of correspondence analysis to analysis using correlation models and association models. For the general canonical correlation model, $M = 3$ and $(\lambda_1^2, \lambda_2^2, \lambda_3^2) = (0.0260, 0.0014, 0.0003)$. The association is rather weak. [Table 10.6](#) contains estimated row and column scores for the correspondence analysis of these three dimensions. Both sets of scores in the first dimension fall in a monotone increasing pattern, except for a slight discrepancy between the first two row scores. This indicates an overall positive association. The scores for the second and third dimensions are close to zero, reflecting the relatively small λ_2 and λ_3 .

Table 10.6 Scores from Correspondence Analysis Applied to [Table 10.5](#)

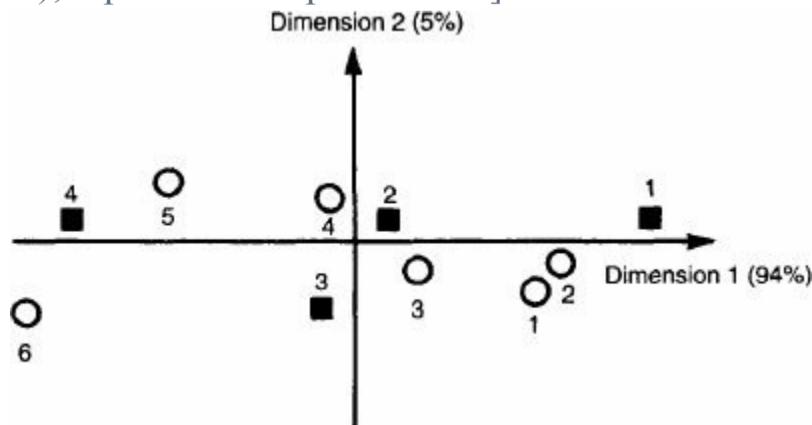
Column Score	Dimension			Row Score	Dimension		
	1	2	3		1	2	3
1	0.260	0.012	0.023	1	0.181	-0.018	0.028
2	0.030	0.024	-0.019	2	0.185	-0.011	-0.026
3	-0.013	-0.069	-0.002	3	0.059	-0.021	-0.010
4	-0.236	0.019	0.016	4	-0.008	0.042	0.011
				5	-0.164	0.044	-0.009
				6	-0.287	-0.061	0.005

Source: Reprinted with permission from the Institute of Mathematical Statistics, based on Goodman (1985).

[Figure 10.3](#) exhibits the results of the correspondence analysis. The horizontal axis has estimates for the first dimension, and the vertical axis has estimates for the second dimension. Six circular points represent the six rows, with point i giving (x_{i1}, x_{i2}) . Similarly, four square points display the estimates (y_{j1}, y_{j2}) . Both sets of points lie close to the horizontal axis, since the first dimension is more important than the second. Row points that are close together represent rows with similar conditional distributions across the columns. Close column points represent columns with similar conditional distributions across rows.

[Figure 10.3](#) Graphical display of scores from first two dimensions of correspondence analysis.

[Based on Escoufier (1982); reprinted with permission.]



Correspondence analysis is used mainly as a descriptive tool. Goodman (1986) developed inferential methods for it. For [Table 10.5](#), inferential analysis reveals that the first dimension, accounting for 94% of the total squared correlation, is adequate for describing the association. Goodman argued for choosing the unsaturated model employing only one dimension and having

graphics display fitted scores for that dimension alone. Then, correspondence analysis is equivalent to a ML analysis using the one-dimensional correlation model ([\(10.14\)](#)). The estimated scores for that model are $(-1.09, -1.17, -0.37, 0.05, 1.01, 1.80)$ for the rows and $(-1.60, -0.19, 0.09, 1.48)$ for the columns. The model fits well ($G^2 = 2.75$, $df = 8$).

The quality of fit and the estimated scores are similar to those shown in Section 10.5.3 for the *RC* model. More parsimonious correlation models also fit these data well, such as ones using equally spaced scores. All analyses of [Table 10.5](#) have yielded similar conclusions about the association. They all neglect, however, that mental health is a natural response variable. It may be more relevant to use an ordinal logistic model.

Like correlation models, a severe limitation of correspondence analysis (CA) is nontrivial generalization to multiway tables. Greenacre (2007) showed displays of several pairwise associations in a single plot and discussed a *multiple correspondence analysis* that applies CA to a large matrix that contains all the pairwise cross-tabulations for pairs of the set of variables being analyzed.

10.5.6 Model Selection and Score Choice for Ordinal Variables

We have presented several ways to use category orderings in model building. To choose among loglinear models, one approach uses the standard models for guidance. If a standard model fits well, simplify by replacing some parameters with structured terms for ordinal classifications.

Association, correlation, and correspondence analysis models have scores for categories of ordinal variables. Parameter interpretations are simplest for equally spaced scores. With parameter scores, the resulting ML estimates of scores need not be monotone. Constrained versions of the models force monotonicity by maximizing the likelihood subject to order restrictions (Agresti et al. 1987, Bartolucci and Forcina 2002, Ritov and Gilula 1991). Disadvantages exist, however, of treating scores as parameters. The model becomes less parsimonious, tests of effects may be less powerful because of a greater df value, and those tests may not use standard distributions.

10.6 EMPTY CELLS AND SPARSENESS IN MODELING CONTINGENCY TABLES

Sparse contingency tables occur when the sample size n is small or when the number of cells is large relative to n . Sparseness is common in tables with many variables. It can have an impact on loglinear model fitting. The following discussion refers to a generic contingency table and model, with cell counts $\{n_i\}$ and expected frequencies $\{\mu_i\}$ for n observations in N cells.

10.6.1 Empty Cells: Sampling Versus Structural Zeros

Sparse tables usually contain some cells with $n_i = 0$. These *empty cells* are of two types: *sampling zeros* and *structural zeros*. In most cases, even though $n_i = 0$, $\mu_i > 0$. It is possible to have observations in the cell, and $n_i > 0$ with sufficiently large n . Such an empty cell is called a *sampling zero*. The empty cells in [Table 10.1](#) for the student survey are sampling zeros.

An empty cell in which observations are impossible is called a *structural zero*. For such cells $\mu_i = 0$ and necessarily $n_i = 0$ and $\hat{\mu}_i = 0$ regardless of n . For a table that cross-classifies cancer patients on their gender, race, and type of cancer, some cancers (e.g., prostate cancer, ovarian cancer) are gender specific. Thus, certain cells have structural zeros. Contingency tables with structural zeros are called *incomplete tables*.

Sampling zeros are part of the data set. A count of 0 is a permissible outcome for a Poisson or a multinomial variate. It contributes to the likelihood function and model fitting. A structural zero, on the other hand, is not an observation and is not part of the data. Sampling zeros are much more common than structural zeros, and the remaining discussion refers to them.

10.6.2 Existence of Estimates in Loglinear Models

Sampling zeros can affect the existence of finite ML estimates of loglinear model parameters. Haberman (1973b, 1974a), generalizing work by Birch (1963) and Fienberg (1970b), studied this. For cell counts \mathbf{n} with expected values $\boldsymbol{\mu}$, Haberman showed results 1 through 5 for Poisson sampling, and by result 6 they apply also to multinomial sampling.

1. The log-likelihood function is a strictly concave function of $\log \boldsymbol{\mu}$.
2. If a ML estimate of $\boldsymbol{\mu}$ exists, it is unique and satisfies the likelihood equations $\mathbf{X}^T \mathbf{n} = \mathbf{X}^T \hat{\boldsymbol{\mu}}$. Conversely, if $\hat{\boldsymbol{\mu}}$ satisfies the model and also the likelihood equations, it is the ML estimate of $\boldsymbol{\mu}$.
3. If all $n_i > 0$, ML estimates of loglinear model parameters exist.
4. Suppose that ML parameter estimates exist for a loglinear model that equates observed and fitted counts in certain marginal tables. Then those marginal tables have uniformly positive counts.
5. If ML estimates exist for a model M , they also exist for any special case of M .
6. For any loglinear model, the ML estimates $\hat{\boldsymbol{\mu}}$ are identical for multinomial and independent Poisson sampling, and those estimates exist in the same situations.

To illustrate, consider the saturated model. By results 2 and 3, when all $n_i > 0$, the ML estimate of $\boldsymbol{\mu}$ is \mathbf{n} . By result 4, parameter estimates do not exist when any $n_i = 0$. Model parameter estimates are contrasts of $\{\log \mu_i\}$, and since $\hat{\boldsymbol{\mu}} = \mathbf{n}$ for the saturated model, the estimates are finite only when all $n_i > 0$.

For unsaturated models, by results 3 and 4 ML estimates exist when all $n_i > 0$ and do not exist when any count is zero in the set of sufficient marginal tables. Suppose that at least one $n_i = 0$ but the sufficient marginal counts are all positive. For hierarchical loglinear models, Glonek et al. (1988) showed that the positivity of the sufficient counts implies the existence of ML estimates if and only if the model is decomposable, which includes the conditional independence models. Models having all pairs of variables associated, however, are more complex. For model (XY, XZ, YZ) , ML estimates exist when only one $n_i = 0$ but may not exist when at least two cells are empty. For instance, ML estimates do not exist for [Table 10.7](#), even though all sufficient statistics (the two-way marginal totals) are positive. See Exercise 10.36.

Table 10.7 Data for Which ML Estimates Do Not Exist for Model $(XY, XZ, YZ)^a$

Z:		1		2	
X	Y:	1	2	1	2
1		0	*	*	*
2		*	*	*	0

^aCells containing * may contain any positive numbers.

Haberman showed that the supremum of the likelihood function is finite. This motivated him to define *extended ML* estimators of $\boldsymbol{\mu}$. These always exist but may equal 0 and, falling on the boundary, need not have the same properties as regular ML estimators. A sequence of estimates satisfying the model that converges to the extended estimate has log likelihood approaching its supremum. In this extended sense, $\hat{\mu}_i = 0$ is the ML estimate of μ_i for the saturated model when $n_i = 0$ and some loglinear parameter estimates are infinite. Lauritzen (1996) gave related results.

When a sufficient marginal count for a factor equals zero, infinite estimates occur for that term. For instance, when a XY marginal total equals zero, infinite estimates occur among $\{\lambda_{ij}^{XY}\}$ for loglinear models such as (XY, XZ, YZ) . Sometimes, however, not even infinite estimates exist. An example is estimating the log odds ratio when both entries in a column of a 2×2 table equal 0.

A value of ∞ (or $-\infty$) for a ML parameter estimate implies that ML fitted values equal 0 in some cells, and some odds ratio estimates equal ∞ or 0. One potential indicator is when the iterative fitting process does not converge, typically because an estimate keeps increasing from cycle to cycle. Some software, however, is fooled after a certain point in the iterative process by the nearly flat

likelihood. It reports convergence, but because of the very slight curvature of the log likelihood, the estimated standard errors are extremely large and numerically unstable. (Similar behavior occurs for logistic models, as we discussed in Section 6.5.) A danger with sparse data is that you might not realize that a true estimated effect is infinite and, as a consequence, report estimated effects and results of statistical inferences that are invalid.

Many ML analyses are unharmed by empty cells. Even when a parameter estimate is infinite, this is not fatal to data analysis. The profile likelihood confidence interval for the true log odds ratio has one endpoint that is finite. For instance, when $n_{11} = 0$ but other $n_{ij} > 0$ in a 2×2 table, $\log \theta = -\infty$ and a confidence interval has form $(-\infty, U)$ for some finite upper bound U .

10.6.3 Effects of Sparseness on X^2 , G^2 , and Model-Based Tests

Section 3.2.3 discussed the adequacy of chi-squared approximations for tests of independence. Similar remarks apply more generally. Although empty cells and sparse tables need not affect model parameter estimates, they can cause sampling distributions of goodness-of-fit statistics to be far from chi-squared. The true sampling distributions converge to chi-squared as $n \rightarrow \infty$, for a fixed number of cells N . The adequacy of the chi-squared approximation depends both on n and N .

The size of n/N that produces adequate approximations for X^2 tends to decrease as N increases (Koehler and Larntz 1980). For fixed n and N , the chi-squared approximation is better for tests with smaller df. For instance, in testing conditional independence in $I \times J \times K$ tables, $G^2[(XZ, YZ)|(XY, XZ, YZ)]$ [with $\text{df} = (I - 1)(J - 1)$] is closer to chi-squared than $G^2(XZ, YZ)$ [with $\text{df} = K(I - 1)(J - 1)$]. The ordinal test of $H_0: \beta = 0$ with the homogeneous linear-by-linear XY association model (10.9) has $\text{df} = 1$, and behaves even better.

The model-based statistics $G^2(M_0|M_1)$ and $X^2(M_0|M_1)$ depend on the data only through the fitted values, and hence only through minimal sufficient statistics for the more complex model. These statistics have null distributions converging to chi-squared as the expected values of the minimal sufficient statistics grow. For most loglinear models, these sufficient statistics refer to marginal tables. Marginal totals are more nearly normally distributed than are single cell counts. Thus, $G^2(M_0|M_1)$ and $X^2(M_0|M_1)$ converge to their limiting chi-squared distribution more quickly than do $G^2(M_0)$ and $X^2(M_0)$, which depend also on individual cell counts. When $\{\hat{\mu}_i\}$ are small but the sufficient marginal totals for M_1 are mostly in at least the range 5 to 10, the chi-squared approximation is usually adequate for model comparison statistics. Haberman (1977a) provided theoretical justification.

10.6.4 Alternative Sparse Data Asymptotics

When large-sample approximations are inadequate, exact small-sample methods are an alternative. When they are infeasible, it is often possible to approximate exact distributions precisely using Monte Carlo methods.

An alternative approach uses sparse asymptotic approximations that apply when the number of cells N increases as n increases. For this approach, $\{\mu_i\}$ need not increase, as they must do in the usual (fixed N , $n \rightarrow \infty$) large-sample theory. Chi-squared statistics then have approximate normal distributions. See Note 10.9.

10.6.5 Adding Constants to Cells of a Contingency Table

One way to obtain finite estimates of all effects and ensure convergence of fitting algorithms is to add a small constant to cell counts. Some algorithms add $\frac{1}{2}$ to each cell, as Goodman (1964b, 1970, 1971a) recommended for saturated models. An example of the beneficial effect of this for a saturated model is bias reduction for estimating an odds ratio in a 2×2 table (Gart and Zweifel 1967). Adding $\frac{1}{2}$ to each cell before fitting an unsaturated model smooths the data too much, however, often causing havoc with sampling distributions. This operation has too conservative an influence on estimated effects and test statistics. The effect is very severe with a large number of cells.

Even for a saturated model, adding $\frac{1}{2}$ to each cell is not a panacea for all purposes. When the ordinary ML estimate of an odds ratio is ∞ , the estimate after adding $\frac{1}{2}$ to each cell is finite, as is the upper endpoint of any confidence interval. However, unless you prefer a Bayesian approach with prior information, it may be more sensible to use an upper bound of ∞ , since no sample evidence suggests that the odds ratio falls below any given value. (Some confidence interval methods that add constants to the data before using Wald formulas, such as in Exercise 1.25, are intended merely to approximate better intervals such as score-test-based intervals.)

When in doubt about the effect of sparse data, perform a sensitivity analysis. For example, for each possibly influential observation, delete it or move it to another cell to see how results vary with small perturbations to the data. Influence diagnostics for GLMs (Williams 1987) are also useful for this purpose. Often, some associations are not affected by empty cells and give stable results for the various analyses, whereas others that are affected are highly unstable. Use caution in making conclusions about an association if small changes in the data are influential.

Other ways exist to smooth data in a less ad hoc manner than adding arbitrary constants to cells. These include penalized likelihood methods (Section 7.4.5) and Bayesian methods as discussed next.

10.7 BAYESIAN LOGLINEAR MODELING

We've just seen that when data are sparse, ML estimates of some model parameters may be infinite. By contrast, the Bayesian approach merges prior information with the sample data, and usually provides shrinkage by which the posterior mean estimate of a parameter is finite, as are both endpoints of posterior intervals.

10.7.1 Estimating Loglinear Model Parameters in Two-Way Tables

For two-way tables, Lindley (1964) proposed Bayesian inference for association parameters, using a Dirichlet prior distribution and its limiting improper prior for the multinomial. He showed that contrasts of log cell probabilities, such as the log odds ratio, have an approximate normal posterior distribution. Using the same structure, Bloch and Watson (1967) provided improved approximations to the posterior distribution and also considered linear combinations of the cell probabilities.

A disadvantage of a Dirichlet prior distribution is that it does not allow for placing structure on the probabilities, such as corresponding to a loglinear model. We could instead put prior distributions on parameters of a loglinear model. Exchangeability within each set of loglinear parameters may be more sensible than the exchangeability of multinomial probabilities that we get with a Dirichlet prior. For the saturated model for two-way tables, a simple approach treats $\{\lambda_i^X\}$, $\{\lambda_j^Y\}$, and $\{\lambda_{ij}^{XY}\}$ as a priori independent with normal priors. We could recognize ordinality by instead using an association model.

Leonard (1975) used a hierarchical approach with the saturated model. For each of $\{\lambda_i^X\}$, $\{\lambda_j^Y\}$, $\{\lambda_{ij}^{XY}\}$, given a mean μ and variance σ^2 , the first-stage prior takes them to be independent and $N(\mu, \sigma^2)$. At the second stage each normal mean is assumed to have an improper uniform distribution over the real line, and σ^2 is assumed to have an inverse chi-squared distribution. Laird (1978), building on this approach, estimated cell probabilities using an empirical Bayesian approach. This approach replaces σ^2 by the mode of the marginal likelihood, after integrating out the loglinear parameters. The analysis shrinks cell proportion estimates toward the fit of the independence model. As $\sigma \rightarrow \infty$, the estimates converge to the sample proportions; as $\sigma \rightarrow 0$, they converge to the independence estimates, $\{p_{i+} p_{+j}\}$. The fitted values have the same row and column marginal totals as the observed data. She noted that the use of a symmetric Dirichlet prior results in posterior mean estimates for cell probabilities that correspond to sample proportions after adding the same count to each cell, whereas her approach permits considerable variability in the amount added or subtracted from each cell to get the estimates.

10.7.2 Example: Polarized Opinions by Political Party

In recent years there has been increasing polarization in the United States between Democrats and Republicans on a variety of issues, such as whether global warming is occurring and whether the government should take steps to try to slow it, whether abortion should be legal, whether homosexuals should have the right to marry, whether health care should be provided for all Americans, whether George W. Bush's tax cuts for the very wealthy should be rescinded, and even whether Barack Obama was born in the United States. Contingency tables cross-classifying many such variables can be created from results of General Social Surveys at sda.berkeley.edu/GSS using such variable names as PARTYID, GRNTAXES, ABANY, and MARHOMO.

[Table 10.8](#) cross-classifies those who identify as strongly Democratic or strongly Republican by opinion about homosexual marriage, for the 2010 GSS for respondents under the age of 40. Of strong Democrats, 39% strongly agreed that homosexuals should have the right to marry, but 0% of strong Republicans felt this way. In restricting the sample by age and to those with strong party ID, the sample size is small ($n = 83$). We regard $n_{21} = 0$ as a sampling zero.

Table 10.8 Political Party Identification and Opinions About Homosexual Marriage

Party Identification	Homosexuals Should Have Right to Marry					Total
	Strongly Agree	Agree	Neutral	Strongly Disagree	Disagree	
Strong Democrat	23	14	10	7	5	59
Strong Republican	0	6	2	7	9	24

Any of the Bayesian strategies just mentioned would yield a positive posterior probability for the empty cell. Here, we recognize the ordinality of the columns by fitting the linear-by-linear model [\(10.5\)](#) in its uniform association form with $\{v_j = j\}$. With uninformative priors, the posterior distribution suggests a strong association. For example, using independent $N(0, \sigma^2)$ prior distributions for all parameters with $\sigma = 1000$ results in the uniform log local odds ratio parameter β having a posterior mean of 0.849 and a posterior standard deviation of 0.213. The implied fitted odds ratio for the four corner cells is $\exp[4(0.849)] \approx 30$.

To find the posterior probability of classification in the empty cell (conditional on being a strong Democrat or strong Republican), we could take the parametric expression $\mu_{21}/(\sum_i \sum_j \mu_{ij})$ for this probability and integrate it with respect to the joint posterior distribution of the model parameters. More simply, using standard output, we could evaluate the probability at the posterior means of all the parameters. This is not identical but gives similar information. For [Table 10.8](#), this gives a cell fitted value of 1.46, an estimated cell probability of 0.02 and an estimated conditional probability that strong Republicans make the strongly agree response of 0.06.

Results using ML are similar, with $\hat{\beta} = 0.813$ ($SE = 0.207$). The model fits adequately, with $X^2 = 6.26$ ($df = 3$). Bayesian analyses for checking model fit are beyond our scope here. Spiegelhalter et al. (2002) presented a *mean posterior deviance* for checking fit and a *deviance information criterion* for comparing models.

10.7.3 Bayesian Loglinear Modeling of Multidimensional Tables

Knuiman and Speed (1988) generalized Leonard's loglinear modeling approach by considering multiway tables and by taking a multivariate normal prior for all parameters collectively rather than univariate normal priors on individual parameters. This permits separate specification of prior information for different interaction terms. They applied this to unsaturated models, computing the posterior mode and using the curvature of the log posterior at the mode to measure precision. King and Brooks (2001) also specified a multivariate normal prior on the loglinear parameters, which induces a multivariate log-normal prior on the expected cell counts. They derived the parameters of this distribution in an explicit form and stated the corresponding mean and covariances of the cell counts.

We've seen that with ML we can analyze a multinomial loglinear model using a corresponding Poisson loglinear model, before conditioning on the sample size. Forster (2010) found corresponding Bayesian results, also using a multivariate normal prior on the model parameters. He discussed conditions for prior distributions such that marginal inferences are equivalent for Poisson and multinomial models. These essentially allow the hyperparameter governing the overall size of the cell means, which disappears after the conditioning that yields the multinomial model, to have an improper prior. Forster also derived necessary and sufficient conditions for the posterior to then be proper, and he related them to conditions for ML estimates to be finite.

Spiegelhalter and Smith (1982) gave an approximate expression for the Bayes factor for a multinomial loglinear model with an improper prior (uniform for the log probabilities) and showed how it related to the standard chi-squared goodness-of-fit statistic. Raftery (1986) noted that this approximation is indeterminate if any cell is empty but is valid with a Jeffreys prior. He also noted that, with large samples, -2 times the log of this approximate Bayes factor is approximately equivalent to Schwarz's BIC model selection criterion.

10.7.4 Graphical Conditional Independence Models

We've seen in Section 10.1 that loglinear conditional independence structure can be summarized by a graph with vertices for the variables and edges between vertices to represent a conditional association. The cell probabilities can be expressed in terms of marginal and conditional probabilities. Independent Dirichlet prior distributions for them induce independent Dirichlet posterior distributions. O'Hagan and Forster (2004, Chap. 12) showed the usefulness of graphical representations for a variety of Bayesian analyses.

Dawid and Lauritzen (1993) introduced the notion of a probability distribution defined over probability measures on a multivariate space that concentrate on a set of such graphs. A special case includes a *hyper Dirichlet* distribution that is conjugate for multinomial sampling and that implies that certain marginal probabilities have a Dirichlet distribution. Madigan and Raftery (1994) and Madigan and York (1995) used this family for graphical model comparison and for constructing posterior distributions for measures of interest by averaging over relevant models. Madigan and York showed how Bayesian graphical models unify many standard discrete data problems. They proposed a Monte Carlo method for Bayesian model averaging. A disadvantage of this approach is that it applies only for decomposable graphical models.

Massam et al. (2009) presented conjugate priors for the loglinear parameters subject to baseline constraints for multinomial sampling. The induced prior on the cell probabilities is a generalization of the hyper Dirichlet prior to nondecomposable graphical models as well as other hierarchical loglinear models.

NOTES

Section 10.1: Conditional Independence Graphs and Collapsibility

10.1 Graphical models: For expositions on graphical models and their conditional independence structure, see Anderson and Böckenholt (2000), Colombi and Giordano (2012), Dobra (2003), Edwards (2000), Edwards and Kreiner (1983), Gottard et al. (2011), Lauritzen (1996), Madigan and York (1995), Marchetti and Lupparelli (2010), Ravikumar et al. (2010), Wermuth and Lauritzen (1983), and Whittaker (1990). Whittaker (1990, Sec. 12.5) summarized connections with various definitions of collapsibility. Khamis (2011) presented an alternative representation using *multigraphs*, in which the vertices represent the cliques of the model and the edges correspond to variables shared by pairs of cliques. For modeling social networks using graphs, see Anderson et al. (1999) and references therein.

10.2 Perfect tables: Darroch (1962) defined a three-way table as *perfect* if for all i, j, k ,

$$\sum_i \frac{\pi_{ij+} \pi_{i+k}}{\pi_{i++}} = \pi_{+j+} \pi_{++k}, \quad \sum_j \frac{\pi_{+jk} \pi_{ij+}}{\pi_{i++}} = \pi_{i++} \pi_{++k}, \\ \sum_k \frac{\pi_{i+k} \pi_{+jk}}{\pi_{++k}} = \pi_{i++} \pi_{+j+}.$$

For perfect tables, homogeneous association implies that

$$\{\pi_{ijk} = \pi_{ij+} \pi_{i+k} \pi_{+jk} / \pi_{i++} \pi_{+j+} \pi_{++k}\}$$

and conditional odds ratios are identical to marginal odds ratios. Whittemore (1978) used perfect tables to illustrate that for $I \times J \times K$ tables with $K > 2$, conditional and marginal odds ratios can be identical even when no pair of variables is conditionally independent. See also Davis (1986b).

Section 10.2: Model Selection and Comparison

10.3 Model selection: For loglinear model selection, see Aitkin (1979), Benedetti and Brown (1978), Brown (1976), Dahinden et al. (2010), Goodman (1970, 1971a), and Wermuth (1976). When a certain model holds, G^2/df has an asymptotic mean of 1. Goodman (1971a) recommended this index for comparing fits. Smaller values represent better fits.

10.4 Partitioning chi-squared: Kullback et al. (1962) and Lancaster (1951) proposed partitionings of chi-squared statistics in multiway tables. Goodman (1970) and Plackett (1962) noted difficulties with their approaches. Lang (1996b) discussed partitionings for more complex models.

Section 10.4: Modeling Ordinal Associations

10.5 $L \times L, R, C$ models: Birch (1965), Goodman (1979a), and Haberman (1974b) introduced special cases of the linear-by-linear association model. Haberman (1974b) expressed the λ^{XY}_{ij} association term with an expansion in orthogonal polynomials. The row effects and column effects models were developed by Goodman (1979a), Haberman (1974b), and Simon (1974). For more general ordinal loglinear models for multiway tables, see Agresti (2010) and references therein on pp. 180–181, Becker (1989a), Becker and Clogg (1989), and Goodman (1986).

Section 10.5: Generalized Loglinear and Association Models, Correlation Models, and Correspondence Analysis

10.6 RC models: Early articles on the *RC* model include Goodman (1979a, 1981a,b) and Andersen (1980, pp. 210–216), apparently partly motivated by earlier work of G. Rasch (see Andersen 1995). Anderson and Böckenholt (2000), Becker (1989a,b, 1990), Becker and Clogg (1989), Choulakian (1988), de Rooij (2008), Goodman (1985, 1986, 1996), and Wong (2010)

discussed generalizations for multiway tables. Xie (1992) adapted it for comparing mobility tables. Anderson (1984) discussed a related model, called the *stereotype model*, as a special case of a baseline-category logit model using parameter response scores. Anderson and Vermunt (2000) showed that RC and related association models arise when observed variables are conditionally independent given a latent variable that is conditionally normal, given the observed variables. Their work generalizes results in Lauritzen and Wermuth (1989) and discussion by Whittaker of van der Heijden et al. (1989). Anderson and Yu (2007) used association models for item response modeling. de Rooij and Heiser (2005) discussed graphical representations for the $RC(M)$ model. Clogg and Shihadeh (1994) surveyed association models and related correlation models. Wong (2010) and Molenberghs and Verbeke (2005, Chap. 6) surveyed association models, with the latter using global odds ratios in many models.

10.7 Correlation models: Kendall and Stuart (1979, Chap. 33) presented canonical correlation methods for contingency tables. See also Williams (1952), who discussed earlier work by R. A. Fisher and others. Karl Pearson often analyzed tables by assuming an underlying bivariate normal distribution (Section 17.1). For estimating that distribution's correlation, see Becker (1989b), Goodman (1981a,b, 1985), Kendall and Stuart (1979, Chaps. 26 and 33), Lancaster (1969, Chap. X), the Pearson (1904) tetrachoric correlation for 2×2 tables, and the Lancaster and Hamdan (1964) polychoric correlation for $I \times J$ tables. Gilula (1984) related the model to latent class models for a two-way table. Gilula and Haberman (1988) analyzed multiway tables with correlation models by treating explanatory variables as a single variable and response variables as a second variable.

10.8 Correspondence analysis (CA): CA gained popularity in France under the influence of Benzécri (1973). Goodman (1996) attributed its origins to H. O. Hartley, publishing under his German birth name (Hirschfeld, 1935). Greenacre (2007) related it to the singular value decomposition of a matrix. For other discussion, see Choulakian (1988), Escoufier (1982), Goodman (1986, 2000), Greenacre (1988), Michailidis and de Leeuw (1998), van der Heijden and de Leeuw (1985), and van der Heijden et al. (1989). Gabriel (1971) discussed related work on biplots.

Section 10.6: Empty Cells and Sparseness in Modeling Contingency Tables

10.9 Sparse ML: For Monte Carlo approximation of exact small-sample distributions, see Booth and Butler (1999), Forster et al. (1996), and Kim and Agresti (1997). For sparse asymptotics with goodness-of-fit testing of a specified multinomial, Koehler and Larntz (1980) showed that a standardized version of G^2 has an approximate normal distribution. Osius and Rojek (1992) considered this for X^2 and G^2 for multinomials with estimated parameters. Koehler (1986) presented limiting normal distributions for G^2 for use in testing models having direct ML estimates. McCullagh (1986) reviewed ways of handling sparse tables and presented an alternative approximation for G^2 . Zelterman (1987) gave normal approximations for X^2 and proposed an alternative statistic. Morris (1975) showed asymptotic normality for a wide class of functions of multinomial counts including chi-squared statistics, and Cressie and Read (1984) showed similar results for the power divergence statistic (Exercise 1.34). Simonoff (1986) proposed a jackknife estimate of the variance of chi-squared statistics under composite hypotheses with the sparse asymptotic framework. For more on ML estimation when a sparse table has empty cells, see Eriksson et al. (2006) and Fienberg and Rinaldo (2011). When a pattern of empty cells forces certain fitted values for a model to equal 0, this affects the df for testing model fit (Haslett 1990).

Section 10.7: Bayesian Loglinear Modeling

10.10 Bayesian loglinear/association modeling: Dellaportas and Forster (1999) and Ntzoufras

et al. (2000) developed MCMC algorithms for choosing among many loglinear models for a high-dimensional table. Kateri et al. (2005) provided a Bayesian analysis of Goodman's $RC(M)$ model. Fienberg and Makov (1998) applied Bayesian loglinear modeling to issues of confidentiality, accounting for model uncertainty via Bayesian model averaging. Agencies often release multidimensional contingency tables that are ostensibly confidential, but the confidentiality can be broken if an individual is uniquely identifiable from the data presentation. For related work in terms of ML estimates, see Dobra et al. (2009).

EXERCISES

Applications

10.1 Refer to the models having fits summarized in [Table 10.2](#).

- a. Which of these models are graphical loglinear models? Why?
- b. Explain why the model symbolized by (ACM, ACR, AMG) is a decomposable, graphical loglinear model. Show that it fits well, with $G^2 = 14.78$ ($df = 16$). Interpret the fit. (Thanks to Giovanni Marchetti for pointing this out.)

10.2 Use [Table 9.3](#) to illustrate the odds ratio collapsibility conditions.

- a. For model (A, C, M) , all conditional odds ratios equal 1.0. Explain why all reported marginal odds ratios equal 1.0.
- b. For model (AC, M) , explain why (i) all conditional odds ratios are the same as the marginal odds ratios, and (ii) all $\hat{\mu}_{ac+} = n_{ac+}$.
- c. For model (AM, CM) , explain why (i) the AC conditional odds ratios of 1.0 need not be the same as the AC marginal odds ratio, and (ii) the AM and CM conditional odds ratios are the same as the marginal odds ratios and all $\hat{\mu}_{a+m} = n_{a+m}$ and $\hat{\mu}_{+cm} = n_{+cm}$.
- d. For model (AC, AM, CM) , explain why (i) no conditional odds ratios need be the same as the related marginal odds ratios, and (ii) the fitted marginal odds ratios must equal the sample marginal odds ratios.

10.3 Refer to the collapsibility condition in Section 10.1.4. For loglinear model (WX, WY, WZ) , what is the impact of collapsing over X , on the other associations? Contrast that with what the conditions suggest, treating group $C = \{X\}$, (i) if $A = \{Z\}$ and $B = \{W, Y\}$, and (ii) if $A = \{Y, Z\}$ and $B = \{W\}$. This shows that different groupings for that condition can give different information.

10.4 [Table 10.9](#) summarizes a study with variables age of mother (A), length of gestation (G) in days, infant survival (I), and number of cigarettes smoked per day during the prenatal period (S). Treat G and I as response variables and A and S as explanatory.

[Table 10.9](#) Data on Gestation and Infant Survival for Exercise 10.4

Age	Smoking	Gestation	Infant Survival	
			No	Yes
<30	<5	≤260	50	315
		>260	24	4012
	5+	≤260	9	40
		>260	6	459
30+	<5	≤260	41	147
		>260	14	1594
	5+	≤260	4	11
		>260	1	124

Source: N. Wermuth, pp. 279–295 in *Proc. 9th International Biometrics Conference*, Vol. 1 (1976). Reprinted with permission from the Biometric Society.

- a. Explain why a loglinear model should include the λ^{AS} term.
- b. Fit the models $(AGIS)$, (AGI, AIS, AGS, GIS) , (AG, AI, AS, GI, GS, IS) , and (AS, G, I) . Identify a subset of models nested between two of these that may fit well.
- c. Use (i) forward selection and (ii) backward elimination to build a model between two of the models listed in (b). Compare the results of the strategies, and interpret the models chosen.

10.5 Consider loglinear model selection for [Table 6.3](#) on P = premarital sex, E = extramarital sex, M = marital status, and G = gender.

- a. Why is it not sensible to consider models that omit the λ^{GM} term?
- b. Using forward selection starting with (GM, E, P) , show that model (GM, GP, EG, EMP) seems reasonable.

c. Using backward elimination, show that (GM, GP, EMP) or (GM, GP, EG, EMP) seems reasonable.

d. Show that the estimated EMP interaction suggests that the effect of extramarital sex on divorce is greater for subjects who had no premarital sex.

e. Use residuals to describe the lack of fit of model (GM, EMP) .

10.6 Refer to the model building in Section 10.2.2. Fit model (7) in [Table 10.2](#). Explain how to interpret the effects of race and gender on these responses.

10.7 For model (AC, AM, CM) with [Table 9.3](#), the standardized residual in each cell equals ± 0.63 . Interpret, and explain why each one has the same absolute value. By contrast, model (AM, CM) has standardized residual ± 3.70 in each cell where $M = \text{yes}$ (e.g., $+3.70$ when $A = C = \text{yes}$) and ± 12.80 in each cell where $M = \text{no}$ (e.g., $+12.80$ when $A = C = \text{yes}$). Interpret.

10.8 For [Table 9.8](#) on auto injuries, conduct a residual analysis with the model of no three-factor interaction to describe the nature of the interaction.

10.9 Refer to the data in Exercise 3.15 on income and job satisfaction.

a. Perform a residual analysis for the independence model. Explain why it suggests that the linear-by-linear association model may fit better. Fit it, compare to the independence model, and interpret.

b. Using standardized scores, find and interpret β .

10.10 For [Table 10.8](#) on opinions about homosexual marriage, fit the linear-by-linear association model and interpret.

10.11 A weak local association may be substantively important for nonlocal categories. Illustrate with the $L \times L$ model for [Table 10.5](#) on mental impairment and parents' SES, showing how the estimated odd ratio for the four corner cells compares to the estimated local odds ratio.

10.12 Refer to [Table 10.3](#) on birth control and premarital sex.

a. Fit the column effects model, using equally spaced row scores. Compare estimated column scores to the equal-interval scores for the $L \times L$ model. Test that the true column scores are equal-interval, given that the model holds. Interpret.

b. The nature of the row categories suggests a special case of the row effects model with the spacing between rows 1 and 2 the same as between rows 3 and 4. Fit this model with equally spaced column scores, test its goodness of fit, and interpret parameter estimates.

10.13 Refer to the previous exercise. Fit the RC model. Interpret the estimated scores. Does it fit significantly better than the uniform association model?

10.14 Replicate the results in Section 10.5.5 for the correlation and correspondence models with [Table 10.5](#) on mental impairment and parents' SES.

10.15 Analyze [Table 10.5](#) on mental impairment using ordinal logistic models. Interpret, and discuss advantages/disadvantages compared with using association models, correlation models, and correspondence analysis.

10.16 Download data from the 2010 General Social Survey at sda.berkeley.edu/GSS cross-classifying opinion about paying higher taxes to help the environment (variable GRNTAXES) and political party identification (PARTYID). Conduct Bayesian inference for an association model. Report posterior mean estimates and their standard deviations for relevant terms describing the associations.

10.17 Conduct a Bayesian loglinear analysis for the GSS data in Exercise 9.6 on attitudes toward abortion, environment, and political party ID. Interpret results, including posterior intervals for conditional odds ratios of interest.

Theory and Methods

10.18 Suppose loglinear model (XY, XZ, YZ) holds. Find $\log \mu_{ij+}$ and explain why marginal associations need not equal conditional associations for this model.

10.19 Consider loglinear model (WX, XY, YZ) . Explain why W and Z are independent given X alone or given Y alone or given both X and Y . When are W and Y conditionally independent?

10.20 For a four-way table, is the WX conditional association the same as the WX marginal association for the loglinear model **(a)** (WX, XYZ) ? **(b)** (WX, WZ, XY, YZ) ? Why?

10.21 Loglinear model M_0 is a special case of loglinear model M_1 .

a. Explain why the fitted values for the two models are identical in the sufficient marginal distributions for M_0 .

b. Haberman (1974a) showed that when $\{\hat{\mu}_i\}$ satisfy any model that is a special case of M_0 , $\sum_i \hat{\mu}_{1i} \log \hat{\mu}_i = \sum_i \hat{\mu}_{0i} \log \hat{\mu}$. In particular, we can regard $\hat{\mu}_0$ as the orthogonal projection of $\hat{\mu}_1$ onto the linear manifold of $\{\log \mu\}$ satisfying M_0 . Using this, show that $G^2(M_0) - G^2(M_1) = 2 \sum_i \hat{\mu}_{1i} \log(\hat{\mu}_{1i}/\hat{\mu}_{0i})$.

10.22 For a three-way table, show that the fit of mutual independence satisfies

$$G^2[(X, Y, Z)] = G^2[(X, Z)] + G^2[(Y, Z)] + G^2[(XZ, YZ)],$$

where models (X, Z) and (Y, Z) refer to the two-way marginal tables (Cheng et al. 2010).

10.23 For T categorical variables X_1, \dots, X_T , explain why:

a. $G^2(X_1, X_2, \dots, X_T) = G^2(X_1, X_2) + G^2(X_1 X_2, X_3) + \dots + G^2(X_1 X_2 \cdots X_{T-1}, X_T)$.

b. $G^2(X_1 \cdots X_{T-1}, X_T) = G^2(X_1, X_T) + G^2(X_1 X_T, X_2 X_3) + \dots + G^2(X_1 X_2 \cdots X_{T-1}, X_1 X_2 \cdots X_{T-2} X_T)$.

10.24 For $I \times 2$ contingency tables, explain why the linear-by-linear association model is equivalent to the linear logit model (5.5) and the column effects model, whereas the row effects model is equivalent to the saturated model.

10.25 Lehmann (1966) defined (X, Y) to be *positively likelihood-ratio dependent* if their joint density satisfies $f(x_1, y_1)f(x_2, y_2) \geq f(x_1, y_2)f(x_2, y_1)$ whenever $x_1 < x_2$ and $y_1 < y_2$. Then, the conditional distribution of $Y|X$ stochastically increases as X (Y) increases (Goodman 1981a).

a. For the $L \times L$ model, show that the conditional distributions of Y and of X are stochastically ordered. What is its nature if $\beta > 0$?

b. In row effects model (10.7), if $\mu_i > \mu_h$, show that the conditional distribution of Y is stochastically higher in row i than in row h .

10.26 Yule (1906) defined a table to be *isotropic* if an ordering of rows and of columns exists such that the local log odds ratios are all nonnegative [see also Goodman (1981a)].

a. Show that a table is isotropic if it satisfies (i) the linear-by-linear association model, (ii) the row effects model, and (iii) the RC model.

b. Explain why a table that is isotropic for a certain ordering is still isotropic when adjacent rows or columns are combined.

10.27 Consider the log likelihood for the linear-by-linear association model.

a. Differentiating with respect to β and evaluating at $\beta = 0$ and null estimates of parameters, show that the score function is proportional to

$$\sum_i \sum_j u_i v_j (p_{ij} - p_{i+} p_{+j}).$$

b. Use the delta method to show that this sum has null SE of

$$\left\{ \left[\sum_i u_i^2 p_{i+} - \left(\sum_i u_i p_{i+} \right)^2 \right] \left[\sum_j v_j^2 p_{+j} - \left(\sum_j v_j p_{+j} \right)^2 \right] / n \right\}^{1/2}.$$

c. Construct a score statistic for testing independence. Show that it is essentially the correlation test (3.16). [Hirotsu (1982) presented a family of score tests for ordered alternatives.]

10.28 For the row effects model (10.7), show that minimal sufficient statistics are $\{n_{i+}\}$, $\{n_{+j}\}$, and $\{\sum_j v_j n_{ij}, i = 1, \dots, I\}$, and show that the likelihood equations equate these to their expected values.

10.29 Show that the column effects model corresponds to a baseline-category logit model for Y

that is linear in scores for X , with slope depending on the paired response categories.

10.30 Refer to the homogeneous linear-by-linear XY association model (10.9).

a. Find the likelihood equations and show that they imply that the fitted marginal XY correlation equals the XY correlation for the sample data.

b. Find the additional likelihood equations that apply for the heterogeneous linear-by-linear XY association model. Explain why, in each stratum, the fitted XY correlation equals the sample correlation.

10.31 When model (XY, XZ, YZ) is inadequate and variables are ordinal, useful models are nested between it and (XYZ) . For ordered scores $\{u_i\}$, $\{v_j\}$, and $\{w_k\}$, consider

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \beta u_i v_j w_k.$$

Define $\theta_{ijk} = \theta_{ij(k+1)}/\theta_{ij(k)} = \theta_{i(j+1)k}/\theta_{i(j)k} = \theta_{(i+1)jk}/\theta_{(i)jk}$. For unit-spaced scores, show there is *uniform interaction*, $\log \theta_{ijk} = \beta$ (Goodman 1979a). Show that log odds ratios for any two variables change linearly across levels of the third variable.

10.32 For a 3×3 table cross-classifying two ordinal variables, show that the model specifying a uniform global odds ratio is a special case of the generalized loglinear model (10.10), by constructing the C , A , and X matrices. Explain why its residual df = 3.

10.33 Explain why the RC model requires scale constraints for the scores. Show that the residual df = $(I - 2)(J - 2)$. Find and interpret the likelihood equations. Explain why the fit is invariant to category orderings.

10.34 For the correlation model (10.14), show (Goodman 1985, 1986):

a. λ is the correlation between the scores.

b. $\sum_i \mu_i (\pi_{ij}/\pi_{+j}) = \lambda v_j$ and $\sum_j v_j (\pi_{ij}/\pi_{i+}) = \lambda \mu_i$. Interpret.

c. With λ close to zero, $\log(\pi_{ij})$ has form $\gamma_i + \delta_j + \lambda \mu_i v_j + o(\lambda)$, where $o(\lambda)/\lambda \rightarrow 0$ as $\lambda \rightarrow 0$. Thus, when the association is weak, the correlation model is similar to the linear-by-linear association model with $\beta = \lambda$ and scores $\{u_i = \mu_i\}$ and $\{v_j = v_j\}$.

10.35 For the general canonical correlation model with M components, show that

$$\sum_{k=1}^M \lambda_k^2 = \sum_i \sum_j (\pi_{ij} - \pi_{i+}\pi_{+j})^2 / \pi_{i+}\pi_{+j}.$$

Thus, the squared correlations partition a dependence measure that is the noncentrality (6.11) of χ^2 for the independence model with $n = 1$. [Goodman (1986) stated other partitionings.]

10.36 Show that ML estimates do not exist for Table 10.7. [Hint: Haberman (1973b, 1974a, p. 398) noted that if $\mu_{111} = c > 0$, then marginal constraints that the model satisfy imply that $\mu_{222} = -c$.]

10.37 For a loglinear model, explain heuristically why the ML estimate of a parameter is infinite when its sufficient statistic takes its maximum or minimum possible value, for given values of other sufficient statistics.

¹Such as the *gnm* function in R developed by Turner and Firth (2007).

CHAPTER 11

Models for Matched Pairs

We next introduce methods for comparing categorical responses for two samples when each observation in one sample pairs with an observation in the other sample. *Such matched-pairs* data commonly occur in studies with repeated measurement of subjects, such as *longitudinal studies* that observe subjects over time on the same categorical scale. Because of the matching, the responses in the two samples are statistically dependent. This is the first of four chapters on methods for observations that are clustered in some way so that it is not sensible to treat them as independent.

[Table 11.1](#) illustrates matched-pairs data. In the 2010 General Social Survey, subjects were asked who they voted for in the 2004 and 2008 Presidential elections. Between these elections, the overall voting population swung in the Democrat direction, going from the Republican George W. Bush being elected in 2004 to the Democrat Barack Obama being elected in 2008. Was there a shift in this direction both for females and for males, and if so, were the shifts of similar magnitude? [Table 11.1](#) shows results for males, for those sampled who voted Democrat or Republican in each election.

Table 11.1 Presidential Votes in 2004 and in 2008, for Males Sampled in 2010 by the General Social Survey

2004 Election	2008 Election		
	Democrat	Republican	Total
Democrat	175	16	191
Republican	54	188	242
Total	229	204	433

Of the 433 males cross-classified in this table, 175 voted Democrat in both elections, 188 voted Republican in both, and 70 changed parties with their votes. The two cells with identical row and column response (the main diagonal of the table) contain most of the sample, since relatively few people changed parties. A strong association exists between the responses, the sample odds ratio being $(175 \times 188)/(16 \times 54) = 38.1$.

For matched pairs with a categorical response, a two-way contingency table with the same row and column categories summarizes the data. The table is *square*. In this chapter we introduce methods for analyzing square tables. In Section 11.1 we describe methods for comparing proportions with a binary response, and Section 11.2 presents logistic regression analyses of such data. For multicategory responses in square tables, Section 11.3 presents nominal and ordinal logistic models for comparing the response distributions, and Section 11.4 introduces loglinear models. In Sections 11.5 and 11.6 we discuss two matched-pairs applications for which models for square tables are useful: analyzing agreement between two observers who rate a common set of subjects, and ranking treatments based on pairwise evaluations.

Section 11.7 extends the models for square tables that result from matched pairs to multiway tables that result from matched sets of observations. In Chapter 12 we extend them further to incorporate explanatory variables.

11.1 COMPARING DEPENDENT PROPORTIONS

For a subject or matched pair randomly selected from the population of interest, let π_{ab} denote the probability of outcome a for the first observation and outcome b for the second. Let n_{ab} count the number of such pairs in a sample of n matched pairs, with $p_{ab} = n_{ab}/n$ the sample proportion. We treat $\{n_{ab}\}$ as a sample from a multinomial $(n; \{\pi_{ab}\})$ distribution. Then, p_{a+} is the proportion in category a for observation 1, and p_{+a} is the corresponding proportion for observation 2. We compare samples by comparing marginal proportions $\{p_{a+}\}$ with $\{p_{+a}\}$. With matched samples, these proportions are correlated, and methods for independent samples are inappropriate.

In this section we consider binary outcomes. When $\pi_{1+} = \pi_{+1}$, then $\pi_{2+} = \pi_{+2}$ also, and there is *marginal homogeneity*. Since

$$\pi_{1+} - \pi_{+1} = (\pi_{11} + \pi_{12}) - (\pi_{11} + \pi_{21}) = \pi_{12} - \pi_{21},$$

marginal homogeneity in 2×2 tables is equivalent to $\pi_{12} = \pi_{21}$. The table then shows *symmetry* across the main diagonal.

11.1.1 Confidence Intervals Comparing Dependent Proportions

One comparison of the marginal distributions uses $\delta = \pi_{+1} - \pi_{1+}$. Let

$$d = p_{+1} - p_{1+} = p_{2+} - p_{+2}.$$

From formula (1.3) for multinomial covariances, $\text{cov}(p_{+1}, p_{1+}) = \text{cov}(p_{11} + p_{21}, p_{11} + p_{12})$ simplifies to $(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})/n$. Thus,

$$(11.1) \quad \text{var}(\sqrt{n} d) = \pi_{1+}(1 - \pi_{1+}) + \pi_{+1}(1 - \pi_{+1}) - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21}).$$

For large samples, d has approximately a normal sampling distribution. A Wald confidence interval for $\delta = \pi_{+1} - \pi_{1+}$ is

$$d \pm z_{\alpha/2} \hat{\sigma}(d),$$

where

$$(11.2) \quad \begin{aligned} \hat{\sigma}^2(d) &= [p_{1+}(1 - p_{1+}) + p_{+1}(1 - p_{+1}) - 2(p_{11}p_{22} - p_{12}p_{21})]/n \\ &= [(p_{12} + p_{21}) - (p_{12} - p_{21})^2]/n, \end{aligned}$$

with the second formula following after substitution and some algebra. Inverting the score test of $H_0: \delta = \delta_0$ is more complex but provides an interval having coverage probability closer to the nominal level (Tango 1998, Tang et al. 2005), as does adding $\frac{1}{2}$ to each cell before computing d and $\hat{\sigma}(d)$ for the Wald interval (Agresti and Min 2005b).

11.1.2 McNemar Test Comparing Dependent Proportions

The hypothesis of marginal homogeneity, $H_0: \pi_{1+} = \pi_{+1}$, is $H_0: \delta = 0$. Under H_0 , an estimated variance of d is

$$(11.3) \quad \hat{\sigma}_0^2(d) = \frac{p_{12} + p_{21}}{n} = \frac{n_{12} + n_{21}}{n^2}.$$

The score test statistic $z_0 = d/\hat{\sigma}_0(d)$ simplifies to

$$(11.4) \quad z_0 = \frac{n_{21} - n_{12}}{\sqrt{n_{21} + n_{12}}}.$$

The square of z_0 is a large-sample chi-squared statistic with $df = 1$. The test using it is called *McNemar's test* (McNemar 1947).

The McNemar statistic depends only on cases classified in *different* categories for the two observations. The $n_{11} + n_{22}$ on the main diagonal are irrelevant to inference about whether π_{1+} and π_{+1} differ. However, *all* cases contribute to inference about *how much* π_{1+} and π_{+1} differ: for instance, to estimating δ and the standard error. Thus, although relatively large values of n_{11} and n_{22} , for given n_{11} and n_{22} , do not give any information about whether there is marginal homogeneity, they do suggest that whatever heterogeneity may exist is small. In summary, pairs with identical outcomes affect the estimated size of the difference between the marginal proportions and its standard error but do not affect statistical significance in terms of whether a nonzero difference truly exists. [Agresti and Min (2003) discussed this issue.]

11.1.3 Example: Changes in Presidential Election Voting

For [Table 11.1](#), the sample proportions of males voting Democrat were $p_{1+} = 191/433 = 0.441$ in 2004 and $p_{+1} = 229/433 = 0.529$ in 2008. Using [\(11.2\)](#), a 95% confidence interval for $\pi_{+1} - \pi_{1+}$ is $0.088 \pm 1.96(0.0189)$, or $(0.051, 0.125)$. We infer that the population percentage of males voting Democrat increased by between about 5% and 13%.

For testing marginal homogeneity, the test statistic [\(11.4\)](#) using the null variance is

$$z_0 = \frac{n_{21} - n_{12}}{\sqrt{n_{21} + n_{12}}} = \frac{54 - 16}{\sqrt{54 + 16}} = 4.54,$$

and the McNemar statistic $z^2_0 = 20.63$ with $df = 1$. The two-sided P -value is 0.000006, extremely strong evidence of a shift in the Democrat direction.

For the corresponding data for females, shown in Exercise 11.2, $p_{1+} = 0.477$ and $p_{+1} = 0.615$. The 95% confidence interval for $\pi_{+1} - \pi_{1+}$ is $0.138 \pm 1.96(0.0167)$, or $(0.106, 0.171)$. The shift toward Democrat seems as if it may be greater for females than males. We can check this with a 95% confidence interval for the difference of differences. Since the females and males are independent samples, this is

$$(0.138 - 0.088) \pm 1.96\sqrt{(0.0189)^2 + (0.0167)^2}, \quad \text{or} \quad (0.001, 0.100).$$

There is evidence of a greater shift for females than males, as much as 10% greater.

11.1.4 Increased Precision with Dependent Samples

The final term of formula (11.1) for $\text{var}(\sqrt{n}d)$ is $-2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})$ is $-2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})$. Based on $\text{cov}(p_{+1}, p_{1+})$, this reflects the dependence between the marginal proportions. By contrast, for *independent* samples of size n each to estimate binomial probabilities π_1 and π_2 , the covariance for the sample proportions is zero, and

$$\text{var}[\sqrt{n} (\text{difference of sample proportions})] = \pi_1(1 - \pi_1) + \pi_2(1 - \pi_2).$$

Dependent samples usually exhibit a positive dependence, with $\log \theta = \log[\pi_{11}\pi_{22}/\pi_{12}\pi_{21}] > 0$; that is $\pi_{11}\pi_{22} > \pi_{12}\pi_{21}$. jri27T2i] From (11.1), positive dependence implies that $\text{var}(d)$ is smaller than when the samples are independent.

A study design using dependent samples can help improve the precision of statistical inferences for within-subject effects. The improvement is substantial when samples are highly correlated. To illustrate, Table 11.1 with dependent samples of size 433 each has a standard error of 0.019 for $d = 0.529 - 0.441$. The two observations have strong association, the sample odds ratio being 38.1. *Independent* samples of size 433 each with $\hat{\pi}_1 - \hat{\pi}_2 = 0.529 - 0.441$ have a standard error of 0.034, nearly twice as large.

11.1.5 Small-Sample Test Comparing Dependent Proportions

The null hypothesis of marginal homogeneity for binary matched pairs is, equivalently, $H_0: \pi_{12} = \pi_{21}$ or $\pi_{21}/(\pi_{21} + \pi_{12}) = 0.50$. For small samples, an exact test conditions on $n^* = n_{21} + n_{12}$. Under H_0 , n_{21} has a binomial $(n^*, \frac{1}{2})$ distribution, for which $E(n_{21}) = \frac{1}{2}n^*$. The P -value for the test uses binomial tail probabilities.

For instance, in analyzing changes in Presidential voting, suppose we focused on those of age less than 30 at the time of the 2004 election. For the 63 males, the counts are $n_{11} = 32$, $n_{12} = 4$, $n_{21} = 8$, and $n_{22} = 19$. Then, $n^* = 8 + 4 = 12$ switched party votes, and the reference distribution is $\text{bin}(12, \frac{1}{2})$. The two-sided P -value is the probability of at least 8 successes or at most 4 successes out of 12 trials; that is,

$$P(n_{21} \geq 8) + P(n_{21} \leq 4) = 2P(n_{21} \geq 8) = 0.388.$$

When $n^* > 10$, the reference binomial distribution is approximately normal with mean $\frac{1}{2}n^*$ and variance $n^*(\frac{1}{2})(\frac{1}{2})$. The standardized normal test statistic equals

$$z = \frac{n_{21} - \frac{1}{2}n^*}{\sqrt{n^*(\frac{1}{2})(\frac{1}{2})}} = \frac{n_{21} - n_{12}}{\sqrt{n_{21} + n_{12}}}.$$

This is identical to the standard normal form (11.4) of the McNemar statistic. With $n_{12} = 4$ and $n_{21} = 8$, $z = 1.15$ has a two-sided P -value of 0.248. The P -value from the large-sample analysis tends to be closer to the binomial two-sided mid P -value. Here, the mid P -value is $2[\frac{1}{2}P(n_{21} = 8) + P(n_{21} \geq 9)] = 0.267$.

11.1.6 Connection Between McNemar and Cochran–Mantel–Haenszel Tests

An alternative representation of binary responses for n matched pairs presents the data in n partial tables, one 2×2 table for each pair. It has columns that are the two possible outcomes for each measurement. Row 1 shows the outcome of the first observation, and row 2 shows the outcome of the second.

[Table 11.2](#) shows the four possible partial tables in this representation. For [Table 11.1](#), the full three-way table has 433 partial tables; 175 look like the one for subject 1 (i.e., Democrat vote in each election), 188 who voted Republican in each election have tables like the one for subject 2, 54 have tables like the one for subject 3, and 16 have tables like the one for subject 4. The 433 subjects from [Table 11.1](#) provide 866 observations in a $2 \times 2 \times 433$ contingency table. Collapsing this table over the 433 partial tables yields a 2×2 table with first row equal to (191, 242) and second row equal to (229, 204). These are the total number of (Democrat, Republican) responses for the two elections. They form the marginal counts in [Table 11.1](#).

Table 11.2 Representation of Four Types of Matched Pairs Contributing to Counts in [Table 11.1](#)

Subject	Election	Vote Response	
		Democrat	Republican
1	2004	1	0
	2008	1	0
2	2004	0	1
	2008	0	1
3	2004	0	1
	2008	1	0
4	2004	1	0
	2008	0	1

For each subject, suppose that the probability of voting Democrat is identical at each election. Then, conditional independence exists between the vote choice and the election date, controlling for subject. The probability of voting Democrat is then also the same for each election in the marginal table collapsed over the subjects. But this implies that the true probabilities for [Table 11.1](#) satisfy marginal homogeneity. Thus, a test of conditional independence in the $2 \times 2 \times 433$ table provides a test of marginal homogeneity for [Table 11.1](#).

To test conditional independence in this three-way table, we can use the Cochran–Mantel–Haenszel (CMH) statistic [\(6.6\)](#). The result of that chi-squared statistic is algebraically identical to the chi-squared form of McNemar’s statistic, namely, $(n_{21} - n_{12})^2 / (n_{12} + n_{21})$ for tables of form [\(11.1\)](#). McNemar’s test is a special case of the CMH test applied to the binary responses of n matched pairs displayed in n partial tables. This connection is not helpful for computational purposes, since the McNemar statistic is simple. But it does suggest ways of handling more complex matched data. With several outcome categories or several observations, we can test marginal homogeneity by applying the generalized CMH tests (Section 8.4) using a single stratum for each subject, with each row representing a particular observation (Darroch 1981; Mantel and Byar 1978).

11.1.7 Subject-Specific and Population-Averaged (Marginal) Tables

We refer to the $2 \times 2 \times n$ table representation of matched-pairs data as the *subject-specific* table. We refer to the 2×2 table of form of [Table 11.1](#) as the *population-averaged* table, since its margins provide direct estimates of population marginal proportions. We'll use the *subject-specific* and *population-averaged* (or *marginal*) terminology in future sections also to refer to models that apply to these two data forms.

11.2 CONDITIONAL LOGISTIC REGRESSION FOR BINARY MATCHED PAIRS

The analyses of Section 11.1 can be expressed in the context of models. Let (Y_{i1}, Y_{i2}) denote the pair of observations for subject (matched-pair) i in the sample, where a “1” outcome denotes category 1 (success) and “0” denotes category 2. Let $P(Y_t = 1)$ denote the mean of $P(Y_{it} = 1)$ for all subjects in the population, where we regard Y_t as the response for a subject randomly selected for observation t . The difference $\delta = P(Y_2 = 1) - P(Y_1 = 1)$ between marginal probabilities occurs as a parameter in the model

$$(11.5) \quad P(Y_t = 1) = \alpha + \delta x_t,$$

where $x_1 = 0$ and $x_2 = 1$; then, $P(Y_1 = 1) = \alpha$ and $P(Y_2 = 1) = \alpha + \delta$. Alternatively, the logit link yields

$$(11.6) \quad \text{logit}[P(Y_t = 1)] = \alpha + \beta x_t.$$

The parameter β is a log odds ratio for the marginal distributions.

11.2.1 Subject-Specific Versus Marginal Models for Matched Pairs

Models (11.5) and (11.6) describe the marginal distributions of responses for the two observations. They are called *marginal models*. For instance, in terms of the population-averaged table, model (11.6) is saturated, and the ML estimate of β is the log odds ratio of marginal proportions, $\hat{\beta} = \log[(p_{+1}/p_{+2})/(p_{1+}/p_{2+})]$. Exercise 11.31 shows its asymptotic variance.

An alternative modeling approach focuses on the subject-specific tables of the form shown in [Table 11.2](#). A model for these tables can allow probabilities to vary by subject, using

$$(11.7) \text{ link}[P(Y_{it} = 1)] = \alpha_i + \beta x_t$$

with subject-specific intercepts. This is called a *subject-specific model*, since the effect β is defined conditional on the subject. Its estimate describes conditional association for the three-way table stratified by subject. By contrast, the effects in marginal models (11.5) and (11.6) are *population-averaged*, since they refer to averaging over the entire population.

The effects in these two types of model can be quite different. To illustrate, for [Table 11.1](#) on Presidential votes in 2004 and in 2008, the ML estimate of the population-averaged effect β in logistic model (11.6) is $\log[(n_{+1}/n_{+2})/(n_{1+}/n_{2+})] = \log[(229/204)/(191/242)] = 0.35$. By contrast, from a result shown below in (11.10), the estimate of the subject-specific effect β in model (11.7) with logit link is $\log(n_{21}/n_{12}) = \log(54/16) = 1.22$. In Section 13.2.3 we'll see why these effects can differ so much.

For the identity link, subject-specific and population-averaged effects are identical. For instance, for the subject-specific model (11.7) with identity link, $\beta = P(Y_{i2} = 1) - P(Y_{i1} = 1)$ for all i , and averaging this over subjects in the population equates β to the δ parameter in model (11.5). For nonlinear link functions, however, the effects differ, as we'll see next.

11.2.2 Logistic Models with Subject-Specific Probabilities

Subject-specific model (11.7) differs from models in earlier chapters by permitting subjects to have their own probability distributions. Cox (1958b, 1970) and Rasch (1961) presented this model with logit link. This model for observation t for subject i is

$$(11.8) \text{logit}[P(Y_{it} = 1)] = \alpha_i + \beta x_t,$$

where $x_1 = 0$ and $x_2 = 1$. That is,

$$P(Y_{i1} = 1) = \frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)}, \quad P(Y_{i2} = 1) = \frac{\exp(\alpha_i + \beta)}{1 + \exp(\alpha_i + \beta)}.$$

The average of $\exp(\alpha_i + \beta x_t)/[1 + \exp(\alpha_i + \beta x_t)]$ for the population does not have the form $\exp(\alpha_i + \beta x_t)/[1 + \exp(\alpha_i + \beta x_t)]$ corresponding to the marginal logistic model (11.6).

Although permitting subject-specific distributions, model (11.8) assumes a common effect β . For subject i , the parameter β compares the response distributions. For each subject, the odds of success for observation 2 are $\exp(\beta)$ times the odds for observation 1.

Given the parameters, with model (11.8) the standard approach assumes independence of responses for different subjects and for the two observations on the same subject. However, averaged over all subjects, the responses are nonnegatively associated. Suppose that $|\beta|$ is small compared with $|\alpha_i|$. A subject with a large positive α_i has high $P(Y_{it} = 1)$ for each t and is likely to have a success each time; a subject with a large negative α_i has low $P(Y_{it} = 1)$ for each t and is likely to have a failure each time. The greater the variability in $\{\alpha_i\}$, the greater the overall positive association between responses, successes (failures) for observation 1 tending to occur with successes (failures) for observation 2. This is true for any β . The positive association reflects the shared value of α_i for each observation in a pair. No association occurs only when $\{\alpha_i\}$ are identical. Thus, the model does account for the dependence in matched pairs. Fitting it takes into account nonnegative association through the structure of the model.

For this model, the large number of $\{\alpha_i\}$ causes difficulties with the fitting process and with the properties of ordinary ML estimators (Exercise 11.29). The remedy of conditional ML treats them as nuisance parameters and maximizes the likelihood function for a conditional distribution that eliminates them. A note on terminology: Model (11.8) is sometimes referred to as a *conditional* model, meaning that its effect β is subject-specific, conditional on the subject. The analyses described below for such models are examples of *conditional* logistic regression; but here the term *conditional* refers to the ML analysis that is performed conditional on sufficient statistics for nuisance parameters, to eliminate those parameters from the likelihood. We introduced this approach in Section 7.3.

11.2.3 Conditional ML Inference for Binary Matched Pairs

For model (11.8), assuming independence of responses for different subjects and for the two observations on the same subject, the joint probability mass function for $\{(y_{i1}, y_{i2}), \dots, (y_{n1}, y_{n2})\}$ is

$$\prod_{i=1}^n \left[\frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)} \right]^{y_{i1}} \left[\frac{1}{1 + \exp(\alpha_i)} \right]^{1-y_{i1}} \left[\frac{\exp(\alpha_i + \beta)}{1 + \exp(\alpha_i + \beta)} \right]^{y_{i2}} \left[\frac{1}{1 + \exp(\alpha_i + \beta)} \right]^{1-y_{i2}}.$$

In terms of the data, this is proportional to

$$\exp \left[\sum_i \alpha_i (y_{i1} + y_{i2}) + \beta \left(\sum_i y_{i2} \right) \right].$$

To eliminate $\{\alpha_i\}$, we condition on their sufficient statistics, the pairwise success totals $[S_i = y_{i1} + y_{i2}]$. Given $S_i = 0$, $P(Y_{i1} = Y_{i2} = 0) = 1$, and given $S_i = 2$, $P(Y_{i1} = Y_{i2} = 1) = 1$. The distribution of (Y_{i1}, Y_{i2}) depends on β only when $S_i = 1$; that is, only when outcomes differ for the two responses. Given $y_{i1} + y_{i2} = 1$, the conditional distribution is

$$\begin{aligned} P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2} | S_i = 1) &= P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}) / [P(Y_{i1} = 1, Y_{i2} = 0) + P(Y_{i1} = 0, Y_{i2} = 1)] \\ &= \frac{\left[\frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)} \right]^{y_{i1}} \left[\frac{1}{1 + \exp(\alpha_i)} \right]^{1-y_{i1}} \left[\frac{\exp(\alpha_i + \beta)}{1 + \exp(\alpha_i + \beta)} \right]^{y_{i2}} \left[\frac{1}{1 + \exp(\alpha_i + \beta)} \right]^{1-y_{i2}}}{\frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)} \frac{1}{1 + \exp(\alpha_i + \beta)} + \frac{1}{1 + \exp(\alpha_i)} \frac{\exp(\alpha_i + \beta)}{1 + \exp(\alpha_i + \beta)}} \\ &= \exp(\beta) / [1 + \exp(\beta)], \quad y_{i1} = 0, \quad y_{i2} = 1 \\ &= 1 / [1 + \exp(\beta)], \quad y_{i1} = 1, \quad y_{i2} = 0. \end{aligned}$$

Again, let $\{n_{ab}\}$ denote the counts for the four possible sequences. For subjects having $S_i = 1$, $\Sigma_i y_{i1} = n_{12}$, the number of subjects having success for observation 1 and failure for observation 2. Similarly, for those subjects, $\Sigma_i y_{i2} = n_{21}$ and $\Sigma_i S_i = n^* = n_{12} + n_{21}$. Since n_{21} is the sum of n^* independent, identical Bernoulli variates, its conditional distribution is binomial with parameter $\exp(\beta) / [1 + \exp(\beta)]$. For testing marginal homogeneity ($\beta = 0$), the parameter equals $\frac{1}{2}$. In summary, the conditional analysis for the logistic model implies that pairs in which $y_{i1} = y_{i2}$ are irrelevant to inference about β . When this model is realistic, it provides justification for comparing marginal distributions using only the $n_{12} + n_{21}$ pairings having outcomes in different categories at the two observations.

Conditional on $S_i = 1$, the joint distribution of the matched pairs is

$$(11.9) \quad \prod_{S_i=1} \left[\frac{1}{1 + \exp(\beta)} \right]^{y_{i1}} \left[\frac{\exp(\beta)}{1 + \exp(\beta)} \right]^{y_{i2}} = \frac{[\exp(\beta)]^{n_{21}}}{[1 + \exp(\beta)]^{n^*}},$$

where the product refers to all pairs having $S_i = 1$. Differentiating the log of this conditional likelihood and equating to 0 and solving yields the conditional ML estimator of β in model (11.8). You can check that it and its standard error are

$$(11.10) \quad \hat{\beta} = \log \left(\frac{n_{21}}{n_{12}} \right), \quad SE = \sqrt{\frac{1}{n_{21}} + \frac{1}{n_{12}}}.$$

11.2.4 Random Effects in Binary Matched-Pairs Model

An alternative remedy to handling the huge number of nuisance parameters in logistic model (11.8) treats $\{\alpha_i\}$ as *random effects*. This regards $\{\alpha_i\}$ as an unobserved random sample from a probability distribution, usually assumed to be $N(\mu, \sigma^2)$ with unknown μ and σ . It eliminates $\{\alpha_i\}$ by averaging with respect to their distribution, yielding a marginal distribution. The likelihood function then depends on β as well as the $N(\mu, \sigma^2)$ parameters. It has only three parameters and is more manageable. For matched pairs with non-negative sample log odds ratio, this approach also yields $\hat{\beta} = \log(n_{21}/n_{12})$ (Neuhaus et al. 1994). This model is an example of a *generalized linear mixed model*, containing both random effects and the fixed effect β . Its analysis is presented in Chapter 13.

Model (11.8) implies that the true odds ratio for each of the n subject-specific partial tables equals $\exp(\beta)$. In Section 6.4.5 we presented the Mantel–Haenszel estimate of a common odds ratio for several 2×2 tables. In fact, that estimator applied to subject-specific tables of the form shown in Table 11.2 is algebraically identical to n_{21}/n_{12} for marginal tables of the form shown in Table 11.1. (Recall that partial tables with responses in only one column do not contribute to the CMH test or Mantel–Haenszel estimate.) In summary, the Mantel–Haenszel estimate, the conditional ML estimate, and (with nonnegative log odds ratio) the ML estimate for the random effects version of logistic model (11.8) yield $\exp(\hat{\beta}) = n_{21}/n_{12}$.

11.2.5 Conditional Logistic Regression for Matched Case–Control Studies

The two observations (y_{i1}, y_{i2}) in a matched pair need not refer to the same subject. For instance, case–control studies that match a single control with each case yield matched-pairs data. For a binary response Y , each case ($Y = 1$) is matched with a control ($Y = 0$) according to criteria that could affect the response. Subjects in the matched pairs are measured on the predictor variable(s) of interest, X , and the XY association is analyzed.

[Table 11.3](#) illustrates. A case–control study of acute myocardial infarction (MI) among Navajo Indians matched 144 victims of MI according to age and gender with 144 people free of heart disease. Subjects were asked whether they had ever been diagnosed as having diabetes ($x = 0$, no; $x = 1$, yes). [Table 11.3](#) has the same form as [Table 11.1](#) except that the levels of X rather than the levels of Y form the rows and the columns.

Table 11.3 Previous Diagnoses of Diabetes for Myocardial Infarction (MI) Case–Control Pairs

MI Controls	MI Cases		
	Diabetes	No Diabetes	Total
Diabetes	9	16	25
No diabetes	37	82	119
Total	46	98	144

Source: J. Coulehan et al., *Am. J. Public Health* 76: 412–414, 1986.
Reprinted with permission from the American Public Health Association.

We can display the data for each matched case–control pair using a partial table of the form shown in [Table 11.2](#), but reversing the roles of X and Y . The X values have four possible patterns, shown in [Table 11.4](#). There are 37 partial tables of type a, since for 37 pairs the case had diabetes and the control did not, 16 partial tables of type b, 9 of type c, and 82 of type d.

Table 11.4 Possible Case–Control Pairs for [Table 11.3](#)

Diabetes	a		b		c		d	
	Case	Control	Case	Control	Case	Control	Case	Control
Yes	1	0	0	1	1	1	0	0
No	0	1	1	0	0	0	1	1

Now, for subject t in matched pair i , consider the model

$$(11.11) \text{ logit}[P(Y_{it} = 1)] = \alpha_i + \beta x_{it}.$$

The probabilities modeled refer to the distribution of Y given X , but the retrospective study provides information only about the distribution of X given Y . We can estimate the odds ratio $\exp(\beta)$, however, because it refers to the XY odds ratio, which relates to both conditional distributions (Sections 2.2.4 and 5.1.4). Even though this study reverses the roles of X and Y in terms of which is fixed and which is random, the conditional ML estimate of $\exp(\beta)$ is $n_{21}/n_{12} = 37/16 = 2.31$.

11.2.6 Conditional Logistic Regression for Matched Pairs with Multiple Predictors

When the binary response has p predictors for case-control or subject-specific matched pairs, the model generalizes to

$$(11.12) \text{logit}[P(Y_{it} = 1)] = \alpha_i + \beta_1 x_{1it} + \beta_2 x_{2it} + \cdots + \beta_p x_{pit},$$

where x_{hit} denotes the value of predictor h for observation t in pair i , $t = 1, 2$. Typically, one predictor is an explanatory variable of interest, such as diabetes status. The others are covariates used to adjust effects, in addition to those already controlled by virtue of using them to form the matched pairs. The conditional ML approach to estimating $\{\beta_j\}$ conditions on sufficient statistics for α_i to eliminate them from the likelihood.

Let $\mathbf{x}_{it} = (x_{1it}, \dots, x_{pit})^T$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$. A generalization of the derivation in Section 11.2.3 shows that

$$(11.13) \begin{aligned} P(Y_{i1} = 0, Y_{i2} = 1 | S_i = 1) &= \exp(\mathbf{x}_{i2}^T \boldsymbol{\beta}) / [\exp(\mathbf{x}_{i1}^T \boldsymbol{\beta}) + \exp(\mathbf{x}_{i2}^T \boldsymbol{\beta})], \\ P(Y_{i1} = 1, Y_{i2} = 0 | S_i = 1) &= \exp(\mathbf{x}_{i1}^T \boldsymbol{\beta}) / [\exp(\mathbf{x}_{i1}^T \boldsymbol{\beta}) + \exp(\mathbf{x}_{i2}^T \boldsymbol{\beta})]. \end{aligned}$$

Dividing numerator and denominator by $\exp(\mathbf{x}_{i1}^T \boldsymbol{\beta})$ shows that the first equation has the form of logistic regression with no intercept and with predictor values $\mathbf{x}_i^* = \mathbf{x}_{i2} - \mathbf{x}_{i1}$. In fact, to obtain conditional ML estimates for model (11.12), we can fit a logistic regression model to these “differing outcome” pairs alone, using artificial response $y^* = 1$ when $(y_{i1} = 0, y_{i2} = 1)$, $y^* = 0$ when $(y_{i1} = 1, y_{i2} = 0)$, no intercept, and predictor values \mathbf{x}_i^* . This addresses the same likelihood as the conditional likelihood (Breslow et al. 1978, Chamberlain 1980).

To illustrate, for model (11.11) with Table 11.3, let $y_i^* = y_{i2} - y_{i1}$ and $x_i^* = \mathbf{x}_{i2} - \mathbf{x}_{i1}$. If $t = 1$ refers to the control and $t = 2$ to the case, then $y_i^* = 1$ always. Since $x_{it} = 1$ represents “yes” for diabetes and $x_{it} = 0$ represents “no,” $(y_i^* = 1, x_i^* = -1)$ for 16 observations, $(y_i^* = 1, x_i^* = 0)$ for $9 + 82 = 91$ observations, and $(y_i^* = 1, x_i^* = +1)$ for 37 observations. The logistic model that forces $\hat{\alpha} = 0$ has $\hat{\beta} = 0.84$. With a single binary predictor, the estimate is identical to $\log(n_{21}/n_{12})$.

11.2.7 Marginal Models and Subject-Specific Models: Extensions

For binary matched-pairs data, Section 11.1 presented analyses for a marginal (i.e., population-averaged) model, and this section presented analyses for a subject-specific model. These models generalize to multinomial responses and to matched sets. For multinomial responses, Chamberlain (1980) proposed conditional ML for matched pairs. For binary responses with a matched set, model (11.12) applies when α_i refers to the set. The matched set might refer to repeated measurements on subject i , or it could refer to a cluster of subjects, such as children from family i or fetuses from litter i .

With extensions of the subject-specific model to matched-set clusters, the conditional ML approach is restricted to estimating β_j that are within-cluster effects, such as occur in case-control and crossover studies. For these, the explanatory variable varies in t for each i . Conditional ML cannot estimate a between-cluster effect. Statistics providing information about such an effect use subject totals at different levels of the relevant explanatory variable; however, those totals sum the sufficient statistics for $\{\alpha_i\}$, so they are themselves fixed and have degenerate distributions after conditioning on the sufficient statistics. An explanatory variable that is constant in t for each i cancels out of the conditional likelihood. [You can observe this for matched pairs with (11.13) for any j for which $x_{ji1} = x_{ji2}$ all i .] For it, at best one can stratify by its levels and fit a model estimating within-cluster effects separately at each level. With subject-specific models, an advantage of using the random effects approach instead of conditional ML is that it is not restricted to estimating within-cluster effects.

In the remainder of this chapter, we emphasize marginal models for matched pairs with multinomial responses. In the following Chapter 12 we deal with marginal model extensions allowing matched sets and explanatory variables. Subject-specific models using a random effects approach have extra computational complexities. We mention briefly some multinomial subject-specific models in this chapter, but we defer most discussion to Chapter 13.

11.3 MARGINAL MODELS FOR SQUARE CONTINGENCY TABLES

Matched-pairs analyses generalize from binary to $I > 2$ outcome categories. A square $I \times I$ table $\{n_{ab}\}$ shows counts of possible sequences (a, b) of outcomes for (Y_{i1}, Y_{i2}) . Let $\pi_{ab} = P(Y_1 = a, Y_2 = b)$ for a randomly selected subject (i.e., the mean of corresponding subject-specific probabilities in the population of interest). Marginal homogeneity is $\pi_{a+} = \pi_{+a}$ for $a = 1, \dots, I$. Marginal models compare $\{\pi_{a+}\}$ and $\{\pi_{+a}\}$.

11.3.1 Marginal Models for Nominal Classifications

For nominal-scale matched-pair responses, marginal model (11.6) for binary matched pairs extends to the baseline-category logit model

$$(11.14) \log[P(Y_t = j)/P(Y_t = I)] = \alpha_j + \beta_j x_t, \quad t = 1, 2, \quad j = 1, \dots, I - 1,$$

where $x_1 = 0$ and $x_2 = 1$. This model has $2(I - 1)$ parameters for the $2(I - 1)$ marginal probabilities. It is saturated.

Marginal homogeneity is the special case $\beta_1 = \dots = \beta_{I-1} = 0$. To fit it, Lipsitz et al. (1990) and Madansky (1963) maximized the multinomial likelihood for $\{n_{ab}\}$ subject to these constraints. Iterative methods produce fitted values $(\hat{\mu}_{ab})$. Comparing these to $\{n_{ab}\}$ using G^2 or X^2 tests marginal homogeneity, with $df = I - 1$.

Bhapkar (1966) tested marginal homogeneity by exploiting the asymptotic normality of marginal proportions. Let $d_a = p_{+a} - p_{a+}$, and let $\mathbf{d}^T = (d_1, \dots, d_{I-1})$. It is redundant to include d_I , since $\sum_a d_a = 0$. The sample covariance matrix $\hat{\psi}$ of $\sqrt{n} \cdot \mathbf{d}$ has elements

$$\hat{\psi}_{ab} = -(p_{ab} + p_{ba}) - (p_{+a} - p_{a+})(p_{+b} - p_{b+}) \quad \text{for } a \neq b,$$

$$\hat{\psi}_{aa} = p_{+a} + p_{a+} - 2p_{aa} - (p_{+a} - p_{a+})^2.$$

Now $\sqrt{n}[\mathbf{d} - E(\mathbf{d})]$ has an asymptotic multivariate normal distribution with estimated covariance matrix $\hat{\psi}$. Under marginal homogeneity, $E(\mathbf{d}) = \mathbf{0}$, and

$$(11.15) W = n\mathbf{d}^T \hat{\psi}^{-1} \mathbf{d}$$

is asymptotically chi-squared with $df = I - 1$. This is a Wald test for parameters in the analog of model (11.14) using the identity link. Stuart (1955) proposed $W_0 = n\mathbf{d}^T \hat{\psi}_0^{-1} \mathbf{d}$, which uses the sample null covariance matrix $\hat{\psi}_0$ and is the score test. This has

$$\hat{\psi}_{ab0} = -(p_{ab} + p_{ba}) \quad \text{for } a \neq b,$$

$$\hat{\psi}_{aa0} = p_{+a} + p_{a+} - 2p_{aa}.$$

Ireland et al. (1969) noted that $W = W_0/(1 - W_0/n)$. For $I = 2$, W_0 is McNemar's statistic, the square of (11.4).

11.3.2 Example: Regional Migration

For the 2010 GSS of American adults. [Table 11.5](#) compares the respondent's region of residence with their region of residence at age 16. Relatively few people changed regions, 84% of the observations falling on the main diagonal. The ML fit of marginal homogeneity, shown in [Table 11.5](#), gives $G^2 = 93.64$ ($df = 3$). Statistics using differences in sample marginal proportions give similar results. For instance, Bhapkar's statistic [\(11.15\)](#) is $W = 90.44$ ($df = 3$).

Table 11.5 Region of Residence in 2010 and at Age 16, with Fit of Marginal Homogeneity Model

Residence at Age 16	Residence in 2010				Total
	Northeast	Midwest	South	West	
Northeast	266 (266)	15 (12.6)	61 (35.7)	28 (16.3)	370 (330.6)
Midwest	10 (12.3)	414 (414)	50 (32.8)	40 (26.2)	514 (485.4)
South	8 (27.7)	22 (46.1)	578 (578)	22 (21.9)	630 (673.6)
West	7 (24.6)	6 (12.7)	27 (27.1)	301 (301)	341 (365.4)
Total	291 (330.6)	457 (485.4)	716 (673.6)	391 (365.4)	1855

Source: 2010 General Social Survey.

The sample marginal percentages for the four regions were (19.9, 27.7, 34.0, 18.4) at age 16 and (15.7, 24.6, 38.6, 21.1) in 2010. The large test statistics reflect the large sample size. To estimate the change for a given region, we apply [\(11.2\)](#) to the collapsed 2×2 table that combines the other regions. A 95% confidence interval for $\pi_{+1} - \pi_{1+}$ is $(0.157 - 0.199) \pm 1.96(0.006)$, or -0.043 ± 0.012 . Similarly, a 95% confidence interval for $\pi_{+2} - \pi_{2+}$ is -0.031 ± 0.013 , for $\pi_{+3} - \pi_{3+}$ is 0.046 ± 0.014 , and for $\pi_{+4} - \pi_{4+}$ is 0.027 ± 0.012 .

11.3.3 Marginal Models for Ordinal Classifications

For ordered categories, marginal model (11.6) for binary matched pairs extends using ordinal logits. With cumulative logits,

$$(11.16) \text{ logit}[P(Y_t \leq j)] = \alpha_j + \beta x_t, \quad t = 1, 2, \quad j = 1, \dots, I - 1,$$

where $x_1 = 0$ and $x_2 = 1$. This model has proportional odds structure (Section 8.2.2). The odds of outcome $Y_2 \leq j$ equal $\exp(\beta)$ times the odds of outcome $Y_1 \leq j$. The model implies stochastically ordered marginal distributions, with $\beta > 0$ meaning that Y_1 tends to be higher than Y_2 . Marginal homogeneity corresponds to $\beta = 0$.

Model fitting treats (Y_1, Y_2) as dependent. The ML approach maximizes the multinomial likelihood for $\{\pi_{ab}\}$. This is not simple. Since the model refers to marginal probabilities $\{P(Y_1 = a) = \pi_{a+}\}$ and $\{P(Y_2 = b) = \pi_{+b}\}$, we cannot substitute the model formula in the kernel $\sum_a \sum_b n_{ab} \log \pi_{ab}$ of the log likelihood, which refers to joint probabilities. We defer discussion of ML model fitting of marginal models to Sections 12.1.4 and 12.1.5. Model (11.16) describes the $2(I - 1)$ marginal probabilities by I parameters, so $\text{df} = I - 2$ for testing fit.

The nominal-scale tests of marginal homogeneity presented in Section 11.3.1 use all $I - 1$ degrees of freedom available for comparisons of I pairs of marginal probabilities. With ordered categories, ordinal tests can focus on a single parameter, $H_0: \beta = 0$, with $\text{df} = 1$. When I is large and the dependence between classifications is strong, the ordinal tests can be much more powerful.

11.3.4 Example: Opinions on Premarital and Extramarital Sex

Refer to [Table 11.6](#). For the 2008 GSS, subjects gave their opinion about premarital sex (a couple having sex before marriage) and extramarital sex (a married person having sex with someone other than the marriage partner). The response categories are 1 = always wrong, 2 = almost always wrong, 3 = wrong only sometimes, 4 = not wrong at all.

Table 11.6 Opinions on Premarital Sex and Extramarital Sex

Premarital Sex	Extramarital Sex				Total
	1	2	3	4	
1	324	6	1	0	331
2	95	16	1	0	112
3	185	30	17	1	233
4	462	120	51	28	661
Total	1066	172	70	29	1337

Source: 2008 General Social Survey.

The sample cumulative marginal proportions are (0.25, 0.33, 0.51, 1.0) for premarital sex and (0.80, 0.92, 0.98, 1.0) for extramarital sex. Responses on premarital sex tended to be more tolerant than those on extramarital sex. The cumulative logit model [\(11.16\)](#) has $\hat{\beta} = 2.780$ ($SE = 0.079$). There is extremely strong evidence that population responses are more negative on extramarital than on premarital sex. The fit of the marginal homogeneity model has $G^2 = 1092.34$ ($df = 3$), and the fit of the ordinal model [\(11.16\)](#) has $G^2 = 105.14$ ($df = 2$). The ordinal model does not fit well, but it fits much better than the marginal homogeneity model. Models to be considered in Section 11.4.7 fit much better yet.

11.4 SYMMETRY, QUASI-SYMMETRY, AND QUASI-INDEPENDENCE

An alternative analysis of square contingency tables directly models the joint distribution using logistic or loglinear models. Some models have marginal homogeneity as a special case.

An $I \times I$ joint distribution $\{\pi_{ab}\}$ satisfies *symmetry* if

$$(11.17) \quad \pi_{ab} = \pi_{ba} \quad \text{whenever } a \neq b.$$

Under symmetry, $\pi_{a+} = \sum_b \pi_{ab} = \sum_b \pi_{ba} = \pi_{+a}$ for all a , so marginal homogeneity occurs. For $I = 2$, symmetry is equivalent to marginal homogeneity, but for $I > 2$, marginal homogeneity can occur without symmetry.

11.4.1 Symmetry as Logistic and Loglinear Models

When all $\pi_{ab} > 0$, symmetry is a logistic model and a loglinear model. In logistic form, it is trivially

$$\log(\pi_{ab}/\pi_{ba}) = 0 \quad \text{for all } a < b.$$

For expected frequencies $\{\mu_{ab} = n\pi_{ab}\}$, it has the loglinear form

$$(11.18) \quad \log \mu_{ab} = \lambda + \lambda_a + \lambda_b + \lambda_{ab},$$

where all $\lambda_{ab} = \lambda_{ba}$. Both classifications have the same single-factor parameters $\{\lambda_a\}$, so $\log \mu_{ab} = \log \mu_{ba}$.

For Poisson or multinomial cell counts $\{n_{ab}\}$, the likelihood equations are

$$\hat{\mu}_{ab} + \hat{\mu}_{ba} = n_{ab} + n_{ba} \quad \text{for all } a < b \quad \text{and} \quad \hat{\mu}_{aa} = n_{aa} \quad \text{for all } a.$$

The main diagonal has perfect fit. The solution that satisfies symmetry is

$$\hat{\mu}_{ab} = \frac{n_{ab} + n_{ba}}{2} \quad \text{for all } a, b.$$

The logistic symmetry model has residual $df = I(I - 1)/2$. For testing symmetry, Bowker (1948) showed that X^2 simplifies to

$$X^2 = \sum \sum_{a < b} \frac{(n_{ab} - n_{ba})^2}{n_{ab} + n_{ba}}.$$

For $I = 2$ this is the chi-squared form of McNemar's statistic, the square of (11.4). The standardized residuals equal

$$r_{ab} = (n_{ab} - n_{ba}) / \sqrt{n_{ab} + n_{ba}}.$$

Only one residual for each pair of categories is nonredundant, since $r_{ab} = -r_{ba}$. They satisfy $\sum \sum_{a < b} r_{ab}^2 = X^2$.

The symmetry model is very simple. Except for a few specialized applications, such as describing intraobserver agreement for pairs of measurements by an observer, it rarely fits well. When the marginal distributions differ substantially, it necessarily fits poorly.

11.4.2 Quasi-symmetry

To accommodate marginal heterogeneity, we can permit the main-effect terms in the symmetry model (11.18) to differ. The resulting loglinear model, called *quasi-symmetry*, is

$$(11.19) \log \mu_{ab} = \lambda + \lambda_a^X + \lambda_b^Y + \lambda_{ab},$$

where $\lambda_{ab} = \lambda_{ba}$ for all $a < b$ (Caußinus 1966). Symmetry is the special case $\lambda_a^X = \lambda_a^Y$ for $a = 1, \dots, I$, and independence is the special case in which all $\lambda_{ab} = 0$. Identifiability requires further constraints, such as $\lambda_I^X = 0$ and all $\lambda_b^Y = 0$ (Exercise 11.36). The likelihood equations for the quasi-symmetry model are

$$\hat{\mu}_{a+} = n_{a+}, \quad a = 1, \dots, I,$$

$$\hat{\mu}_{+b} = n_{+b}, \quad b = 1, \dots, I,$$

$$(11.20) \hat{\mu}_{ab} + \hat{\mu}_{ba} = n_{ab} + n_{ba} \quad \text{for } a \leq b.$$

Only one of the first two sets of equations is needed. The other is redundant, given the other two. The residual df = $(I - 1)(I - 2)/2$. From (11.20), $\hat{\mu}_{aa} = n_{aa}$ for $a = 1, \dots, I$. Otherwise, the likelihood equations do not have a direct solution and are solved using iterative methods such as Newton–Raphson or IPF.

The meaning of quasi-symmetry is less obvious than symmetry. However, it usually fits much better and has greater scope. The main-effect parameters determine the relative sizes of μ_{ab} and μ_{ba} . For example, with constraints that set all $\lambda_b^Y = 0$, $\log(\mu_{ab}/\mu_{ba}) = \lambda_a^X - \lambda_b^X$.

The quasi-symmetry model has multiplicative form

$$(11.21) \pi_{ab} = \alpha_a \beta_a \gamma_{ab}, \quad \text{where } \gamma_{ab} = \gamma_{ba} \quad \text{all } a < b$$

and all parameters are positive. The symmetry model is (11.21) with $\alpha_a = \beta_a$ for all a . This equation indicates that a table satisfying quasi-symmetry is the cellwise product of a table satisfying independence with one satisfying symmetry. The association symmetry implies that odds ratios on one side of the main diagonal are identical to corresponding odds ratios on the other side. In fact, the model can be defined by properties such as

$$(11.22) \frac{\mu_{ab} \mu_{II}}{\mu_{aI} \mu_{Ib}} = \frac{\mu_{ba} \mu_{II}}{\mu_{bI} \mu_{Ia}} \quad \text{for all } a < b$$

or $\theta_{ab} = \theta_{ba}$ for all local odds ratios. Goodman (1979a) referred to it as the *symmetric association* model.

Another way to interpret the model parameters relates to subject-specific logistic models. Consider the adaptation of baseline-category logit model (11.14) to a subject-specific model,

$$(11.23) \log[P(Y_{it} = j)/P(Y_{it} = I)] = \alpha_{ij} + \beta_j x_t, \quad t = 1, 2, \quad j = 1, \dots, I - 1,$$

where $x_1 = 1$ and $x_2 = 0$. This has the additive form of binary model (11.8) for each j . The model implies, averaging over subjects, that the quasi-symmetry model (11.19) holds for the $I \times I$ population-averaged table with $\{\beta_j = \lambda_j^X\}$, when we constrain $\lambda_I^X = 0$ and all $\lambda_j^Y = 0$ (Section 12.2.7 shows this in the binary case). In fact, for the conditional ML analysis that conditions out $\{\alpha_{ij}\}$, the conditional ML estimates of $\{\hat{\beta}_j\}$ are identical to the ML estimates of $\{\hat{\lambda}_j^X\}$ for the quasi-symmetry model with these constraints (Conaway 1989).

11.4.3 Marginal Homogeneity and Quasi-symmetry

Marginal homogeneity is not equivalent to a loglinear model. However, quasi-symmetry is a useful model for studying marginal homogeneity. Caussinus (1966) showed that symmetry is equivalent to quasi-symmetry and marginal homogeneity holding simultaneously. We have seen that symmetry implies both quasi-symmetry and marginal homogeneity. Now we give Caussinus's argument for the converse, that the joint occurrence of quasi-symmetry and marginal homogeneity implies symmetry.

From (11.21), if quasi-symmetry holds, $\pi_{ab} = \alpha_a \beta_b \gamma_{ab}$, where $\gamma_{ab} = \gamma_{ba} > 0 > 0$ for all $a < b$. Equivalently,

$$\pi_{ab} = \rho_a \delta_{ab},$$

where $\rho_a = \alpha_a / \beta_a$ and $\delta_{ab} = \beta_a \beta_b \gamma_{ab}$ also satisfies $\delta_{ab} = \delta_{ba} > 0 > 0$ for all $a < b$. If there is also marginal homogeneity, then

$$\pi_{j+} = \rho_j \sum_b \delta_{jb} = \sum_a \rho_a \delta_{aj} = \pi_{+j},$$

or

$$\rho_j = \left(\sum_a \rho_a \delta_{aj} \right) / \left(\sum_b \delta_{jb} \right) = \left(\sum_a \rho_a \delta_{aj} \right) / \left(\sum_b \delta_{bj} \right), \quad j = 1, \dots, I.$$

Thus, each ρ_j is a weighted average of $\{\rho_a\}$, with weights $\{\delta_{aj} / \sum_b \delta_{bj} > 0, a = 1, \dots, I\}$. Any set $\{\rho_a\}$ satisfying this must be identical. Otherwise, there would be a ρ_j that is no greater than any ρ_a but smaller than at least one, and hence it could not be a positive weighted average of all of them. But since $\{\rho_a\}$ are identical, $\pi_{ab} = \rho_a \delta_{ab} = \rho_b \delta_{ab} = \rho_b \delta_{ba} = \pi_{ba}$, so symmetry holds. Thus, a table that satisfies both quasi-symmetry and marginal homogeneity also satisfies symmetry. Since the converse holds,

$$(11.24) \text{ quasi-symmetry + marginal homogeneity} = \text{symmetry}.$$

It follows that when quasi-symmetry (QS) holds, marginal homogeneity (MH) is equivalent to symmetry (S), which is $\{\lambda_a^X = \lambda_a^Y, a = 1, \dots, I\}$ in the QS model. Thus, conditional on quasi-symmetry, testing marginal homogeneity is equivalent to testing symmetry. A test of marginal homogeneity compares fit statistics for the symmetry and quasi-symmetry models,

$$(11.25) G^2(S | QS) = G^2(S) - G^2(QS),$$

with $df = I - 1$. This is an alternative to approaches discussed in Section 11.3.1 using baseline-category logit marginal models.

11.4.4 Quasi-independence

Square tables usually exhibit positive dependence, manifested by larger counts on the main diagonal than the independence model predicts. Conditional on the event that a matched pair falls off the main diagonal, though, the relationship may have a simple structure.

A square contingency table satisfies *quasi-independence* when the variables are independent, given that the row and column outcomes differ. This has the loglinear form

$$\log \mu_{ab} = \lambda + \lambda_a^X + \lambda_b^Y + \delta_a I(a = b), \quad (11.26)$$

where $I(\cdot)$ is the indicator function,

$$I(a = b) = \begin{cases} 1, & a = b \\ 0, & a \neq b. \end{cases}$$

This adds a parameter to the independence model for each cell on the main diagonal. The first three terms on the right-hand side of (11.26) specify independence, and $\{\delta_a\}$ permit $\{\mu_{aa}\}$ to depart from this pattern and have arbitrary positive values. When $\delta_a > 0$, μ_{aa} is larger than under independence.

The likelihood equations for quasi-independence are

$$\hat{\mu}_{a+} = n_{a+}, \quad \hat{\mu}_{+a} = n_{+a}, \quad \hat{\mu}_{aa} = n_{aa}, \quad a = 1, \dots, I.$$

A perfect fit occurs on the main diagonal, but independence holds for the remaining cells. The model implies that odds ratios equal 1.0 for all rectangularly formed 2×2 tables in which all cells fall off the main diagonal. The model can be fitted using Newton–Raphson or IPF. The model has I more parameters than the independence model, so its residual df = $(I - 1)^2 - I$. It applies to tables with $I \geq 3$.

Quasi-independence is the special case of quasi-symmetry (11.21) in which $\{\gamma_{ab} \text{ for } a \neq b\}$ are identical. They are equivalent when $I = 3$ (Causinus 1966, p. 146).

11.4.5 Example: Migration Revisited

For [Table 11.5](#) on migration patterns, not surprisingly the independence model fits terribly ($G^2 = 2828.22$, $df = 9$). The symmetry model is also unpromising. For instance, 40 people moved from the Midwest to the West, but only 6 people made the reverse move. The deviance for testing symmetry is $G^2 = 100.48$ ($df = 6$). The much smaller deviance compared with the independence model reflects that it forces a perfect fit on the main diagonal, where most observations occur.

The quasi-symmetry model has $G^2 = 6.98$, with $df = 3$. [Table 11.7](#) displays its fit. The difference $G^2(S \mid QS) = 100.48 - 6.98 = 93.50$ ($df = 3$) shows extremely strong evidence of marginal heterogeneity. Results are similar to those quoted in Section 11.3.2 for the likelihood-ratio test based on baseline-category logit model [\(11.14\)](#), for which $G^2 = 93.64$ ($df = 3$).

Table 11.7 Fits of Models to Migration [Table 11.5](#)

Residence at Age 16	Residence in 2010				Total
	Northeast	Midwest	South	West	
Northeast	266 (266, 266) ^a	15 (18.3, 15.7)	61 (54.8, 58.5)	28 (30.9, 29.8)	370
Midwest	10 (10.9, 9.3)	414 (414, 414)	50 (57.0, 55.2)	40 (32.1, 35.5)	514
South	8 (9.1, 10.5)	22 (15.9, 16.8)	578 (578, 578)	22 (26.9, 24.8)	630
West	7 (5.0, 5.2)	6 (8.8, 10.5)	27 (26.2, 24.2)	301 (301, 301)	341
Total	291	457	716	391	1855

^aFirst value is quasi-independence fit, second is quasi-symmetry fit; both models give a perfect fit on main diagonal.

The lack of symmetry in cell probabilities reflects the marginal heterogeneity. The effects can be described using the quasi-symmetry model parameter estimates. With the constraints $\lambda_4^X = 0$ and all $\lambda_b^Y = 0$, we have $\{\hat{\lambda}_1^X = 1.74, \hat{\lambda}_2^X = 1.21, \hat{\lambda}_3^X = 0.02\}$. For example, since $\log(\mu_{14}/\mu_{41}) = \lambda_1^X$, the estimated odds of moving from the Northeast to the West were $\exp(1.74) = 5.7$ times the odds of moving from the West to the Northeast.

Quasi-independence states that for people who moved, their region in 2010 is independent of their region at age 16. [Table 11.7](#) also contains its fitted values, for which $G^2 = 9.21$ ($df = 5$). Its fit is not significantly poorer than the quasi-symmetry model.

11.4.6 Ordinal Quasi-symmetry

The loglinear models presented so far for square tables treat classifications as nominal. With ordered categories, more parsimonious models are useful. Let $u_1 \leq \dots \leq u_I$ denote ordered scores for both the row and columns. An *ordinal quasi-symmetry model* is

$$(11.27) \log \mu_{ab} = \lambda + \lambda_a + \lambda_b + \beta u_b + \lambda_{ab},$$

where $\lambda_{ab} = \lambda_{ba}$ for all $a < b$. It is the special case of the quasi-symmetry model (11.19) in which

$$\lambda_b^Y - \lambda_b^X = \beta u_b$$

has a linear trend. Symmetry is the special case $\beta = 0$.

This model has logistic representation,

$$(11.28) \log(\pi_{ab}/\pi_{ba}) = \beta(u_b - u_a) \quad \text{for } a \leq b.$$

This models the logit of the conditional probability of cell (a,b) , given response sequence (a,b) or (b,a) . The greater the value of $|\beta|$, the greater the difference between π_{ab} and π_{ba} and hence between the marginal distributions. Its likelihood equations are

$$\sum_a u_a \hat{\mu}_{a+} = \sum_a u_a n_{a+}, \quad \sum_b u_b \hat{\mu}_{+b} = \sum_b u_b n_{+b},$$

$$\hat{\mu}_{ab} + \hat{\mu}_{ba} = n_{ab} + n_{ba} \quad \text{for } a < b.$$

The fitted marginal counts need not equal the observed marginal counts. However, dividing the first two equations by n shows that they have the same means, for the chosen scores.

When $\beta \neq 0$, this model implies stochastically ordered margins. When $\beta > 0$, responses have a higher mean in the column distribution. Like the ordinal marginal models (Section 11.3.3), this model concentrates the marginal effect on $\text{df} = 1$. A test of marginal homogeneity ($H_0: \beta = 0$) uses

ordinal quasi-symmetry + marginal homogeneity = symmetry.

The likelihood-ratio test statistic compares the deviance for symmetry and ordinal quasi-symmetry.

We can fit model (11.28) with logistic model software: Identify (n_{ab}, n_{ba}) as binomial with $n_{ab} + n_{ba}$ trials, and fit a logistic model with no intercept and predictor $x = u_b - u_a$.

11.4.7 Example: Premarital and Extramarital Sex Revisited

For [Table 11.6](#) on attitudes toward premarital and extramarital sex, a cursory glance at the data reveals that the symmetry model is inadequate ($G^2 = 1243.07$, $df = 6$). By comparison, quasi-symmetry fits well ($G^2 = 0.65$, $df = 3$).

The simpler model of ordinal quasi-symmetry also fits well: With scores (1, 2, 3, 4), $G^2 = 2.81$ ($df = 5$). The ML estimate $\hat{\beta} = -3.035$. From [\(11.28\)](#), the estimated probability that outcome on premarital sex is x categories more positive than the outcome on extramarital sex equals $\exp(3.035x)$ times the reverse probability.

11.5 MEASURING AGREEMENT BETWEEN OBSERVERS

We now discuss an application, analyzing agreement between two observers, that uses models for matched pairs. We illustrate with [Table 11.8](#). This shows ratings by two pathologists, labeled *A* and *B*, who separately classified 118 slides regarding the presence and extent of carcinoma of the uterine cervix. The rating scale has the ordered categories (1) negative, (2) atypical squamous hyperplasia, (3) carcinoma in situ, and (4) squamous or invasive carcinoma.

Table 11.8 Diagnoses of Carcinoma

Pathologist A	Pathologist B ^a				Total
	1	2	3	4	
1	22 (8.5)	2 (-0.5)	2 (-5.9)	0 (-1.8)	26
2	5 (-0.5)	7 (3.2)	14 (-0.5)	0 (-1.8)	26
3	0 (-4.1)	2 (-1.2)	36 (5.5)	0 (-2.3)	38
4	0 (-3.3)	1 (-1.3)	17 (0.3)	10 (5.9)	28
Total	27	12	69	10	118

^aValues in parentheses are standardized residuals for the independence model.
Source: N. Holmquist et al., *Arch. Pathol.* **84**: 334–345, 1967. Reprinted with permission from the American Medical Association. See also Landis and Koch (1977).

11.5.1 Agreement: Departures from Independence

For [Table 11.8](#), let π_{ab} denote the probability that observer A classifies a slide in category a and observer B classifies it in category b . Then π_{aa} is the probability that they both choose category a , and $\sum_a \pi_{aa}$ is the total probability of agreement. Perfect agreement occurs when $\sum_a \pi_{aa} = 1$.

With subjective scales, agreement is less than perfect. Analyses focus on describing strength of agreement and detecting patterns of disagreement. *Agreement* and *association* are distinct facets of the joint distribution. Strong agreement requires strong association, but strong association can exist without strong agreement. If observer A consistently rates subjects one category higher than observer B , strength of agreement is poor even though the association is strong.

Evaluations of agreement can compare $\{n_{ab}\}$ to the values $\{n_{a+b}/n\}$ predicted under independence. That model is a baseline, showing the agreement expected if no association existed between ratings. Normally, it fits poorly if even mild agreement exists, but its cell standardized residuals (Section 3.3.1) show patterns of agreement and disagreement. Ideally, standardized residuals are large positive on the main diagonal and large negative off that diagonal. The sizes are influenced by sample size n , with larger values tending to occur with larger n .

The independence model fits [Table 11.8](#) poorly ($G^2 = 117.96$, df = 9). That table also reports the standardized residuals. The large positive residuals on the main diagonal indicate that agreement for each category is greater than expected by chance, especially for the first category. Off the main diagonal they are primarily negative. Disagreements occurred less than expected under independence, although the evidence of this is weaker for categories closer together. The most common disagreements were observer B choosing category 3 and observer A instead choosing category 2 or 4.

11.5.2 Using Quasi-independence to Analyze Agreement

More complex models add components that relate to agreement beyond that expected under independence. A useful generalization is quasi-independence (11.26), which adds main-diagonal parameters $\{\delta_a\}$. For Table 11.8, this model has $G^2 = 13.18$ (df = 5). It fits much better than independence, but some lack of fit remains. Table 11.9 shows the fit.

Table 11.9 Fitted Values for Carcinoma Diagnoses of Table 11.8

		Pathologist <i>B</i>			
		1	2	3	4
Pathologist <i>A</i>					
1		22 (22, 22) ^a	2 (0.7, 2.4)	2 (3.3, 1.6)	0 (0.0, 0.0)
2		5 (2.4, 4.6)	7 (7, 7)	14 (16.6, 14.4)	0 (0.0, 0.0)
3		0 (0.8, 0.4)	2 (1.2, 1.6)	36 (36, 36)	0 (0.0, 0.0)
4		0 (1.9, 0.0)	1 (3.0, 1.0)	17 (13.1, 17.0)	10 (10, 10)

^aQuasi-independence model fit followed by quasi-symmetry model fit.

Loglinear models can directly address the association component of agreement. For two observations, suppose each observer classifies one in category a and one in category b . The odds that the observers agree rather than disagree on which is in category a and which is in category b equal

$$(11.29) \quad \tau_{ab} = \frac{\pi_{aa}\pi_{bb}}{\pi_{ab}\pi_{ba}} = \frac{\mu_{aa}\mu_{bb}}{\mu_{ab}\mu_{ba}}.$$

Under the quasi-independence loglinear model,

$$\tau_{ab} = \exp(\delta_a + \delta_b).$$

Larger $\{\delta_a\}$ represent stronger agreement and stronger association. For instance, for Table 11.8, $\delta_2 = 0.60$ and $\delta_3 = 1.90$, and $\tau_{23} = 12.3$. The degree of agreement/association also seems quite strong for other pairs of categories.

11.5.3 Quasi-symmetry and Agreement Modeling

For [Table 11.8](#), the quasi-independence model shows some lack of fit. Given that the pathologists disagree, some association remains between ratings. For observer agreement tables, this is common. Quasi-symmetry ([11.19](#)) often fits much better, because it permits association. For [Table 11.8](#), it has $G^2 = 0.98$ ($df = 2$). [Table 11.9](#) displays the fit. It is not unusual for tables to have many empty cells. When $n_{ab} + n_{ba} = 0$ for any pair (such as categories 1 and 4 in [Table 11.8](#)), the ML fitted values for quasi-symmetry in those cells must also be zero since one of its likelihood equations is $\hat{\mu}_{ab} + \hat{\mu}_{ba} = n_{ab} + n_{ba}$. You should eliminate those cells from the fitting process to get the proper residual df value.

Under quasi-symmetry, $t_{ab} = \exp(\hat{\lambda}_{aa} + \hat{\lambda}_{bb} - \hat{\lambda}_{ab} - \hat{\lambda}_{ba})$, where $\hat{\lambda}_{ab} = \hat{\lambda}_{ba}$. For categories 2 and 3 of [Table 11.8](#), for instance, $t_{23} = 10.7$. The model also yields information about similarity of marginal distributions. The simpler symmetry model that forces the margins to be identical fits [Table 11.8](#) poorly ($G^2 = 39.18$, $df = 5$). The statistic $G^2(S|QS) = 39.18 - 0.98 = 38.20$ ($df = 3$) provides strong evidence of marginal heterogeneity. In [Table 11.8](#), differences in marginal proportions are substantial in each category but the first. The marginal heterogeneity is one reason that the agreement is not stronger.

Models for agreement can take ordering of categories into account (Agresti 2010, Sec. 8.5.3). Conditional on observer disagreement, a tendency usually remains for high (low) ratings by one observer to occur with relatively high (low) ratings by the other observer.

11.5.4 Kappa: A Summary Measure of Agreement

An alternative approach summarizes agreement with a single index. For nominal scales, the most popular measure is *Cohen's kappa* (Cohen 1960). It compares the probability of agreement $\sum_a \pi_{aa}$ to that expected if the ratings were independent, $\sum_a \pi_{a+}\pi_{+a}$, by

$$\kappa = \frac{\sum_a \pi_{aa} - \sum_a \pi_{a+}\pi_{+a}}{1 - \sum_a \pi_{a+}\pi_{+a}}.$$

The denominator equals the numerator with $\sum_a \pi_{aa}$ replaced by its maximum possible value of 1, corresponding to perfect agreement. Kappa equals 0 when the agreement merely equals that expected under independence. It equals 1.0 when perfect agreement occurs. The stronger the agreement, the higher is κ , for given marginal distributions. Negative values occur when agreement is weaker than expected by chance, but this rarely happens.

For multinomial sampling, the sample value $\hat{\kappa}$ has a large-sample normal distribution. Its estimated asymptotic variance is

$$\hat{\sigma}^2(\hat{\kappa}) = \frac{1}{n} \left\{ \frac{P_o(1 - P_o)}{(1 - P_e)^2} + \frac{2(1 - P_o)[2P_oP_e - \sum_a p_{aa}(p_{a+} + p_{+a})]}{(1 - P_e)^3} \right. \\ \left. + \frac{(1 - P_o)^2 [\sum_a \sum_b p_{ab}(p_{b+} + p_{+a})^2 - 4P_e^2]}{(1 - P_e)^4} \right\},$$

where $P_o = \sum_a p_{aa}$ and $P_e = \sum_a p_{a+}p_{+a}$ (Fleiss et al. 1969). It is rarely plausible that agreement is no better than expected by chance. Thus, rather than testing $H_0: \kappa = 0$, it is more relevant to estimate strength of agreement by interval estimation of κ .

For [Table 11.8](#), $P_o = 0.636$ and $P_e = 0.281$. Sample kappa equals $(0.636 - 0.281)/(1 - 0.281) = 0.493$. The difference between observed agreement and that expected under independence is about 50% of the maximum possible difference. The estimated standard error is 0.057, so κ apparently falls between about 0.38 and 0.60, moderately strong agreement.

11.5.5 Weighted Kappa: Quantifying Disagreement

Kappa treats classifications as nominal. When categories are ordered, the seriousness of a disagreement depends on the difference between the ratings. For nominal classifications also, some disagreements may be considered more severe than others. The measure *weighted kappa* (Spitzer et al. 1967) uses weights $\{w_{ab}\}$ satisfying $0 \leq w_{ab} \leq 1$, with all $w_{aa} = 1$ and all $w_{ab} = w_{ba}$ to describe closeness of agreement. One possibility is $\{w_{ab} = 1 - |a - b|/(I - 1)\}$, for which agreement is greater for cells nearer the main diagonal. Fleiss and Cohen (1973) suggested $\{w_{ab} = 1 - (a - b)^2/(I - 1)^2\}$. The weighted agreement is $\sum_a \sum_b w_{ab} \pi_{ab}$ and weighted kappa is

$$\kappa_w = \frac{\sum_a \sum_b w_{ab} \pi_{ab} - \sum_a \sum_b w_{ab} \pi_{a+} \pi_{+b}}{1 - \sum_a \sum_b w_{ab} \pi_{a+} \pi_{+b}}.$$

Controversy surrounds the utility of kappa and weighted kappa, partly because their values depend strongly on the marginal distributions. The same diagnostic rating process can yield quite different values, depending on the proportions of cases of the various types (Exercise 11.42). In summarizing a contingency table by a single number, the reduction in information can be severe. An alternative is to find kappa separately for each outcome category, for the 2×2 table in which the other categories are combined. Or, models can provide more detailed description of the agreement and disagreement structure, as we noted in Sections 11.5.2 and 11.5.3.

11.5.6 Extensions to Multiple Observers

With several observers, ordinary loglinear models are not usually relevant. Their description of agreement and association between two observers is conditional on ratings-by the others. It is more relevant to study this marginally, without conditioning on the other ratings. Models for the pairwise agreement and association structure then focus simultaneously on all pairs of two-way marginal distributions (Becker and Agresti 1992).

Other approaches have also been used. For instance, generalizations of kappa summarize pairwise agreements or multiple agreements (Fleiss et al. 2003, Sec. 18.3; Landis and Koch 1977). A mixture model assumes latent classes of subjects for whom the observers agree and subjects for whom they disagree. Such an analysis is shown in Section 14.1.3.

11.6 BRADLEY–TERRY MODEL FOR PAIRED PREFERENCES

Sometimes, categorical outcomes result from pairwise evaluations. A common example is athletic competitions, when the outcome for a team or player consists of categories (win, lose). Another example is pairwise comparison of product brands, such as two brands of wine of some type. When a wine critic rates I brands of New Zealand sauvignon blanc, it might be difficult to establish an outright ranking, especially if I is large. However, for any given pair, the critic could probably state a preference after tasting them at the same occasion. An overall ranking of the wines could then be based on the pairwise preferences. We next present a model for doing this.

11.6.1 Bradley–Terry Model

Bradley and Terry (1952) proposed a logistic model for paired evaluations. Let Π_{ab} denote the probability that a is preferred to b . Suppose that $\Pi_{ab} + \Pi_{ba} = 1$ for all pairs; that is, a tie cannot occur. The Bradley–Terry model is

$$(11.30) \quad \log \frac{\Pi_{ab}}{\Pi_{ba}} = \beta_a - \beta_b.$$

Alternatively,

$$\prod_{ab} = \exp(\beta_a)/[\exp(\beta_a) + \exp(\beta_b)].$$

Thus, $\Pi_{ab} = \frac{1}{2}$ when $\beta_a = \beta_b$ and $\Pi_{ab} > \frac{1}{2}$ when $\beta_a > \beta_b$. Identifiability requires a constraint such as $\beta_I = 0$. Since the model describes all the pairwise probabilities ($\{\Pi_{ab}\}$ for $a < b$) by $(I - 1)$ parameters, residual df = $\binom{I}{2} - (I - 1)$.

For $a < b$, let N_{ab} denote the sample number of evaluations, with a preferred n_{ab} times and b preferred $n_{ba} = N_{ab} - n_{ab}$ times. A square contingency table with empty cells on the main diagonal summarizes results. When the N_{ab} comparisons are independent with probability Π_{ab} for each, n_{ab} has a $\text{bin}(N_{ab}, \Pi_{ab})$ distribution. If evaluations for different pairs are also independent, ordinary methods for logistic models apply for fitting the model.

11.6.2 Example: Major League Baseball Rankings

[Table 11.10](#) shows results of the 2011 season for the five baseball teams in the Eastern Division of the American League. For instance, of games between Boston and New York, Boston won 12 and New York won 6 (one of the few bright points in a disastrous season for the Red Sox). [Table 11.10](#) shows the population of regular-season games. We regard this as a sample estimate of a conceptual distribution representing the long-run performance of teams as constituted in 2011.

Table 11.10 Results of 2011 Season for American League (Eastern Division) Baseball Teams

Winning Team	Losing Team				
	Boston	New York	Tampa Bay	Toronto	Baltimore
Boston	—	12	6	10	10
New York	6	—	9	11	13
Tampa Bay	12	9	—	12	9
Toronto	8	7	6	—	12
Baltimore	8	5	9	6	—

Source: www.baseball-reference.com/leagues/AL/2011-standings.shtml.

Table 11.11 Results of Fitting Bradley–Terry Model to Baseball Data of [Table 11.10](#)

Team	Winning Percentage	$\hat{\beta}_a$	SE
Boston	52.8	0.454	0.304
New York	54.2	0.499	0.305
Tampa Bay	58.3	0.635	0.307
Toronto	45.8	0.229	0.303
Baltimore	38.9	0.000	—

We fitted the Bradley–Terry model as a logistic model for $(\frac{5}{2}) = 10$ independent binomial samples, using an appropriate model matrix and no intercept. The assumption of binomial sampling is suspect, as for any particular pair of teams the probability of victory for a team may vary according to factors not considered here, such as the quality of the starting pitcher for each team. Nonetheless, the model fits adequately ($G^2 = 7.70$, $df = 6$). [Table 11.11](#) displays the sample proportion of games each team won and the model estimates of $\{\hat{\beta}_a\}$, setting $\hat{\beta}_5 = 0$. When Boston played New York, the estimated probability that Boston won is

$$\hat{\Pi}_{12} = \exp(\hat{\beta}_1)/[\exp(\hat{\beta}_1) + \exp(\hat{\beta}_2)] = 0.489.$$

The standard error of each $\hat{\beta}_a - \hat{\beta}_b$ is about 0.30, so not much evidence exists of a difference among these teams. The likelihood-ratio statistic for testing $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$ is 5.48 with $df = 4$ ($P = 0.24$).

11.6.3 Example: Home Team Advantage in Baseball

The analysis above does not recognize which team is the home team. Most sports have a home field advantage: A team is more likely to win when it plays at its home city. [Table 11.12](#) contains results for 2011 according to the (home team, away team) classification. For instance, when Boston was the home team, it beat New York 5 times and lost 4 times; when New York was the home team, it beat Boston 2 times and lost 7 times. Now for all $a \neq b$, let Π_{ab}^* denote the probability that team a beats team b , when a is the home team. Consider logistic model

Table 11.12 Wins/Losses by Home Team and Away Team, for American League (Eastern Division) Baseball Teams in 2011 Season

Home Team	Away Team				
	Boston	New York	Tampa Bay	Toronto	Baltimore
Boston	—	5-4	2-7	6-3	6-3
New York	2-7	—	6-3	7-2	7-2
Tampa Bay	5-4	6-3	—	6-3	3-6
Toronto	2-7	6-3	5-4	—	6-3
Baltimore	5-4	3-6	3-6	3-6	—

Source: www.baseball-reference.com/games.

$$(11.31) \quad \log \frac{\Pi_{ab}^*}{1 - \Pi_{ab}^*} = \alpha + (\beta_a - \beta_b).$$

When $\alpha > 0$, a home field advantage exists.

For [Table 11.12](#), model (11.31) describes 20 binomial distributions with 5 parameters. It has $G^2 = 19.41$ ($df = 15$). The estimate of the home-field parameter is $\hat{\alpha} = 0.080$ with $SE = 0.154$. For two evenly matched teams, the home team had estimated probability $e^{0.080}/[1 + e^{0.080}] = 0.521$ of winning. In fact, of the 180 games between teams in this division in 2010, the home team won 94, or 52.2%.

For this model, the estimated team effects are $\hat{\beta}_1 = 0.453$, $\hat{\beta}_2 = 0.498$, $\hat{\beta}_3 = 0.482$, $\hat{\beta}_4 = 0.379$, and $\hat{\beta}_5 = 0.000$, with SE values of about 0.30 for differences. When Boston played New York, the estimated probability of a Boston win was 0.509 at Boston and 0.469 at New York. Perhaps surprisingly, the simpler model that has $\alpha = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$, corresponding to each game having result comparable to flipping a fair coin, has $G^2 = 23.54$ ($df = 20$) and does not give a significantly poorer fit (change in deviance = 4.13, $df = 5$, $P = 0.53$).

Model (11.31) is a useful generalization of the Bradley–Terry model whenever an *order effect* exists. For instance, in pairwise taste evaluations, the product tasted first may have a slight advantage.

11.6.4 Bradley–Terry Model and Quasi-symmetry

Fienberg and Larntz (1976) showed that the Bradley–Terry model is a logistic formulation of the quasi-symmetry model (11.19). For quasi-symmetry, given that an observation is in cell (a,b) or (b,a) , the logit of the conditional probability of cell (a,b) equals

$$\log \frac{\mu_{ab}}{\mu_{ba}} = (\lambda + \lambda_a^X + \lambda_b^Y + \lambda_{ab}^{XY}) - (\lambda + \lambda_b^X + \lambda_a^Y + \lambda_{ba}^{XY})$$

$$= (\lambda_a^X - \lambda_a^Y) - (\lambda_b^X - \lambda_b^Y) = \beta_a - \beta_b,$$

where $\beta_a = \lambda_a^X - \lambda_a^Y$. Estimates $\{\hat{\lambda}_a^X\}$ and $\{\hat{\lambda}_a^Y\}$ for quasi-symmetry yield $\{\hat{\beta}_a\}$ for the Bradley–Terry model, and recall that we can constrain all $\lambda_a^Y = 0$ for quasi-symmetry.

11.6.5 Extensions to Ties and Ordinal Pairwise Evaluations

The Bradley–Terry model extends to ordinal comparisons, such as the evaluation scale (much better, slightly better, the same, slightly worse, much worse) in comparing two products. With cumulative logits and an I -category evaluation scale, let Y_{ab} denote the response for a comparison of a with b . The model is

$$\text{logit}[P(Y_{ab} \leq j)] = \alpha_j + (\beta_a - \beta_b).$$

Since $P(Y_{ab} \leq j) = P(Y_{ba} > I - j) = 1 - P(Y_{ba} \leq I - j)$, it follows that $\text{logit}[P(Y_{ab} \leq j)] = -\text{logit}[P(Y_{ba} \leq I - j)]$. Thus, necessarily, $\alpha_j = -\alpha_{I-j}$.

The most common ordered preference scale is (win, tie, lose). Then, $\alpha_1 = -\alpha_2$.

11.7 MARGINAL MODELS AND QUASI-SYMMETRY MODELS FOR MATCHED SETS

Methods for matched pairs extend to matched sets. Here we present mainly the loglinear modeling approach, with a brief discussion of marginal models.

11.7.1 Marginal Homogeneity, Complete Symmetry, and Quasi-symmetry

Let (Y_1, Y_2, \dots, Y_T) denote the T responses for a randomly selected matched set. With I response categories, a contingency table with I^T cells summarizes the possible outcomes. Let $\mathbf{j} = (j_1, \dots, j_T)$ denote the cell having $Y_t = j_t$, $t = 1, \dots, T$. Let $\pi_j = P(Y_t = j_t, t = 1, \dots, T)$. Then

$$P(Y_t = j) = \pi_{+ \dots + j + \dots +},$$

where the j subscript is in position t , and $\{P(Y_t = j), j = 1, \dots, I\}$ is the marginal distribution for Y_t .

This T -way table satisfies *marginal homogeneity* if

$$P(Y_1 = j) = P(Y_2 = j) = \dots = P(Y_T = j) \text{ for } j = 1, \dots, I.$$

It satisfies *complete symmetry* if

$$\pi_j = \pi_k$$

for any permutation $\mathbf{k} = (k_1, \dots, k_T)$ of $\mathbf{j} = (j_1, \dots, j_T)$. Complete symmetry implies marginal homogeneity, but the converse does not hold except when $T = I = 2$.

Complete symmetry is a loglinear model. For $\mu_j = n\pi_j$, one representation is

$$\log \mu_j = \lambda_{ab\dots m},$$

where a is the minimum of (j_1, \dots, j_T) , b is the next smallest, \dots , and m is the maximum. In a three-way table, for instance, $\log \mu_{122} = \log \mu_{212} = \log \mu_{221} = \lambda_{122}$. The number of $\{\lambda_{ab\dots m}\}$ parameters is the number of ways of selecting T out of I items with replacement, which is $\binom{I+T-1}{T}$. Thus, residual df = $I^T - (I + T - 1)!/[T!(I - 1)!]$.

An I^T table satisfies *quasi-symmetry* if

$$(11.32) \quad \log \mu_j = \lambda_{1j_1} + \lambda_{2j_2} + \dots + \lambda_{Tj_T} + \lambda_{ab\dots m},$$

where $\lambda_{ab\dots m}$ is defined as in the complete symmetry model. It has symmetric association and higher-order interaction terms, but permits each single-factor marginal distribution to have its own parameters. Identifiability requires constraints such as $\lambda_{tI} = 0$ for each t and equating one set of main-effect terms to 0. This model has $(I - 1)(T - 1)$ more parameters than complete symmetry.

When quasi-symmetry (11.32) holds, marginal homogeneity is equivalent to complete symmetry. The statistic

$$G^2(S|QS) = G^2(S) - G^2(QS)$$

tests marginal homogeneity. Under complete symmetry, it has large-sample chi-squared distribution with df = $(I - 1)(T - 1)$.

11.7.2 Types of Marginal Symmetry

A general type of symmetry for I^T tables has marginal homogeneity and complete symmetry as special cases. For an I^T table, $P(Y_{t_1} = j_1, \dots, Y_{t_h} = j_h)$, where h is between 1 and T , is an h -dimensional marginal probability, $h = 1$ giving single-variable marginal probabilities. There is h th-order marginal symmetry if for all h -tuples $\mathbf{j} = (j_1, \dots, j_h)$, this probability is the same for each permutation of \mathbf{j} and for all combinations $\mathbf{t} = (t_1, \dots, t_h)$ of h of the T responses.

For $h = 1$, first-order marginal symmetry is marginal homogeneity. Second-order marginal symmetry occurs if for all t and u , $P(Y_t = a, Y_u = b)$ is the same and the equality holds for all pairs of outcomes (a, b) . In other words, the two-way marginal tables exhibit symmetry, and they are identical. T th-order marginal symmetry in an I^T table is complete symmetry. When h th-order symmetry holds, i th-order marginal symmetry holds for any $i < h$. For instance, complete symmetry implies second-order marginal symmetry, which itself implies marginal homogeneity.

This hierarchy is mathematically attractive. However, the higher-order symmetries are usually too restrictive to fit well in practice.

11.7.3 Comparing Binary Marginal Distributions in Multiway Tables

Usually, the multivariate dependence structure among repeated responses is of less interest than their marginal distributions. For instance, in treating a chronic condition (such as migraine headaches or a phobia) with some treatment, the primary goal might be to study whether the probability of success increases over the T weeks of a treatment period. The T success probabilities refer to the T first-order marginal distributions. In Sections 11.2.1 and 11.3 we compared marginal distributions for matched pairs ($T = 2$) using models that apply directly to the marginal distributions. We now extend this approach to $T > 2$.

For the binary case, the marginal logistic model (11.6) for matched pairs extends to

$$(11.33) \text{ logit}[P(Y_t = 1)] = \alpha + \beta_t, \quad t = 1, \dots, T,$$

with a constraint such as $\beta_T = 0$ or $\alpha = 0$. This model is saturated, describing T marginal probabilities by T parameters. Marginal homogeneity, for which $P(Y_1 = 1) = \dots = P(Y_T = 1)$, is the special case $\beta_1 = \dots = \beta_T$. Even though this case has only one parameter, ML fitting is not simple. Let $\boldsymbol{\pi}$ denote the vector of the probabilities π_i for the possible \mathbf{i} . They specify the joint distribution of (Y_1, \dots, Y_T) for the 2^T table that cross-classifies the T responses. The multinomial likelihood refers to $\boldsymbol{\pi}$ rather than the T marginal probabilities $\{P(Y_t = 1)\}$. Fitting methods are described in Section 12.1.4.

Let n_i denote the sample cell count in cell \mathbf{i} . The kernel of the log likelihood $L(\boldsymbol{\pi})$ is $\sum_i n_i \log \pi_i$. Let $L(\mathbf{p})$ denote the log likelihood evaluated at the sample proportions $\{p_i = n_i/n\}$, the ML fit of model (11.33). Let $L(\hat{\boldsymbol{\pi}}^{MH})$ denote the maximized log likelihood assuming marginal homogeneity. The likelihood-ratio test of marginal homogeneity (Lipsitz et al. 1990, Madansky 1963) uses

$$(11.34) \quad -2[L(\hat{\boldsymbol{\pi}}^{MH}) - L(\mathbf{p})] = 2 \sum_i n_i \log(p_i / \hat{\pi}_i^{MH}).$$

The asymptotic null chi-squared distribution has $\text{df} = T - 1$.

11.7.4 Example: Attitudes Toward Legalized Abortion

For the GSS data in [Table 11.13](#), subjects indicated whether they support legalized abortion in each of three situations. For this 2^3 table, let μ_{hij} denote the expected frequency for response sequence (h, i, j) for the three situations. Consider the model

Table 11.13 Support for Legalizing Abortion in Three Situations

	Sequence of Responses on the Three Items ^a							
	(1,1,1)	(1,1,2)	(2,1,1)	(2,1,2)	(1,2,1)	(1,2,2)	(2,2,1)	(2,2,2)
Counts	466	3	39	1	71	3	423	147
QS fit	466	0.4	40.2	2.4	72.4	4.2	420.4	147

^aItems (1) family has very low income and cannot afford more children, (2) woman is not married and does not want to marry the man, (3) woman's health is seriously endangered by the pregnancy. 1, yes; 2, no.

Source: Data from 2010 General Social Survey.

$$\log \mu_{hij} = \lambda_{abc},$$

with interaction term λ_{111} when $(h, i, j) = (1,1,1)$, λ_{112} when $(h, i, j) = (1,1,2)$ or $(1,2,1)$ or $(2,1,1)$, λ_{122} when $(h, i, j) = (1,2,2)$ or $(2,1,2)$ or $(2,2,1)$, and λ_{222} when $(h, i, j) = (2,2,2)$. This model implies a complete symmetry pattern of probabilities. Its fit has $G^2 = 965.94$ with $df = 4$. The lack of fit is not surprising, as, for instance, $n_{212} = 1$ whereas $n_{221} = 423$.

Adding main-effect terms for the three situations provides the quasi-symmetry model. It fits much better, having $G^2 = 8.30$ with $df = 2$. [Table 11.13](#) shows its fitted values. Its estimated main-effect terms $(-4.60, -5.18, 0)$ show greater support for legalized abortion when a woman's health is seriously endangered than in the other two cases. The slight lack of fit reflects that the joint distribution has departures from a symmetric association structure. For example, the loglinear model with only two-factor association terms, which has $G^2 = 0.33$ with $df = 1$, has fitted log odds ratios of 4.30 for situations 1 and 2, 1.95 for situations 1 and 3, and 2.20 for situations 2 and 3. There is a stronger conditional association between opinions on the two non-health-related situations.

We can test marginal homogeneity by the likelihood-ratio statistic comparing symmetry and quasi-symmetry models, $965.94 - 8.30 = 957.64$, with $df = 2$. The likelihood-ratio statistic [\(11.34\)](#) for directly testing marginal homogeneity is 647.54, with $df = 2$. Either statistic gives extremely strong evidence of marginal heterogeneity. For simultaneous confidence intervals comparing proportions in support of legalization for pairs of situations, with overall error probability ≤ 0.05 , the Bonferroni method uses confidence coefficient $(1 - 0.05/3) = 0.9833$ for each. From formula [\(11.1\)](#), the estimate $0.029 = 0.471 - 0.441$ of the difference between situations (1) and (2) has an estimated standard error of 0.0092. The Wald confidence interval for the true difference is $0.029 \pm 2.39(0.0092)$, or $(0.01, 0.05)$. The intervals are $(0.36, 0.43)$ comparing (3) and (1) and $(0.39, 0.46)$ comparing (3) and (2). We infer that there are differences of opinion between each pair of situations, with very large differences between health (3) and the other two.

11.7.5 Marginal Homogeneity for a Multicategory Response

The binary marginal model (11.33) extends to multinomial responses. With baseline-category logits for I outcome categories, the saturated model is

$$(11.35) \log[P(Y_t = j)/P(Y_t = I)] = \beta_{tj}, \quad t = 1, \dots, T, \quad j = 1, \dots, I - 1.$$

Marginal homogeneity, whereby $P(Y_1 = j) = \dots = P(Y_T = j)$ for $j = 1, \dots, I - 1$, is the special case in which

$$\beta_{1j} = \beta_{2j} = \dots = \beta_{Tj}, \quad j = 1, \dots, I - 1.$$

The likelihood-ratio test of marginal homogeneity comparing the two models has form (11.34) and $\text{df} = (T - 1)(I - 1)$.

For an ordinal response, an unsaturated model that is more complex than marginal homogeneity focuses on location shifts among the T margins. One such model is

$$(11.36) \text{logit}[P(Y_t \leq j)] = \alpha_j + \beta_t, \quad t = 1, \dots, T, \quad j = 1, \dots, I - 1,$$

with constraint such as $\beta_T = 0$. This model can be fitted using ML methodology presented in Section 12.1.4. Marginal homogeneity is the special case $\beta_1 = \dots = \beta_T$, tests of which have $\text{df} = T - 1$.

11.7.6 Wald and Generalized CMH Score Tests of Marginal Homogeneity

Let $p_j(t)$ denote the sample proportion in category j for response Y_t , let

$$\bar{p}_j = \sum_t p_j(t)/T, \quad d_j(t) = p_j(t) - \bar{p}_j,$$

and let \mathbf{d} denote the vector of $\{d_j(t), t = 1, \dots, T - 1, j = 1, \dots, I - 1\}$. Let $\hat{\Psi}$ denote the estimated covariance matrix of $\sqrt{n} \mathbf{d}$. Bhapkar (1973) proposed the Wald statistic

$$(11.37) \quad W = n \mathbf{d}^T \hat{\Psi}^{-1} \mathbf{d}$$

for the general alternative. This generalizes (11.15) and has a large-sample chi-squared distribution with $\text{df} = (I - 1)(T - 1)$.

Other statistics are special cases of the generalized multicategory Cochran–Mantel–Haenszel (CMH) statistic of Section 8.4.3. Recall that for the binary case ($T = 2$) with matched pairs ($T = 2$), the CMH statistic applies to a three-way table (see, e.g., Table 11.2) in which each stratum shows the two outcomes for a given subject. A generalization of Table 11.2 provides n strata of $T \times I$ tables. The k th stratum gives the T outcomes for subject k . Row t in a stratum has a 1 in the column that is the outcome for observation t , and 0 in all other columns (or 0 in every column if that observation is missing). Probability distributions for the subject-stratified setup naturally relate to subject-specific models such as logistic model (11.23), rather than to marginal models. However, conditional independence in this three-way table (given subject) corresponds to an exchangeability among variables in the I^T table that implies marginal homogeneity. A generalized CMH test of conditional independence in the $T \times I \times n$ table also tests marginal homogeneity using a sampling distribution generated under the stronger exchangeability condition (Darroch 1981). For an ordinal response with fixed scores, the generalized CMH statistic for detecting variability among T means is appropriate.

When $I = 2$ and $T = 2$, this CMH approach is equivalent to McNemar's statistic. When $I = 2$ but $T > 2$, the generalized CMH statistic treating the T responses as unordered is identical to a statistic proposed by Cochran (1950, Exercise 11.49), often referred to as *Cochran's Q*.

In the next chapter we present marginal models in more general contexts. We extend the analyses of this chapter to incorporate matched sets with explanatory variables.

NOTES

Section 11.1: Comparing Dependent Proportions

11.1 McNemar generalized: Altham (2010), Copas (1973), Gart (1969), Kenward and Jones (1994), and Miettinen (1969) considered generalizations of matched-pairs designs. Altham (2010) discussed alternatives to McNemar's test when each response is missing some observations. Miettinen generalized the McNemar test to case-control sets having several controls per case. The [Table 11.2](#) representation is then useful. Each of n matched sets forms a stratum of a $2 \times 2 \times n$ table with one observation in column 1 (the case) and several observations in column 2 (the controls). Westfall et al. (2010) proposed multiple comparison methods for multiple McNemar tests with dependent or independent samples. Lloyd (2008) surveyed exact methods and proposed an unconditional method based on maximizing the P -value when using an estimate of the nuisance parameter. Altham (1971), Consonni and La Rocca (2008), and Ghosh et al. (2000) presented Bayesian analyses for binary matched pairs. For some of these and some unconditional approaches, inferences about marginal homogeneity also use the main-diagonal observations (Liang and Zeger 1988, Suissa and Shuster 1991).

Section 11.4: Symmetry, Quasi-symmetry, and Quasi-independence

11.2 Quasi-symmetry: For other discussion of quasi-symmetry (QS), see Darroch (1981) and McCullagh (1982). It contains as a special case other useful models, such as the ones in Sections 11.4.4 and 11.6.4. Kateri and Papaioannou (1997) showed that under certain conditions QS is the closest model to complete symmetry in terms of Kullback–Leibler distance. Gottard et al. (2011) proposed graphical models with colored edges to represent QS structure among multivariate responses having some common categorical scales. Results at the end of Section 11.4.2 relating conditional ML to QS extend to multiple occasions using a multivariate form ([11.32](#)) of QS (Agresti 1997, Bhapkar and Darroch 1990, Conaway 1989, Darroch 1981, Tjur 1982; see also Section 14.2.7). The effect β in ordinal QS ([11.27](#)) relates to the main effect in a subject-specific adjacent-categories-logit model (Agresti 1993). The symmetry model generalizes in other ways for ordinal responses, such as with the conditional symmetry model

$$\log(\pi_{ab}/\pi_{ba}) = \tau, \quad a < b$$

(Bishop et al. 1975, pp. 285–286), which implies that for all $a < b$,

$$P(Y_1 = a, Y_2 = b | Y_1 < Y_2) = P(Y_1 = b, Y_2 = a | Y_1 > Y_2).$$

For other generalizations, see Agresti (2010, Sec. 8.3), Goodman (1979b, 1985), Hout et al. (1987), and Kateri and Agresti (2007).

11.3 Quasi-independence: The term *quasi-independence* originated in Goodman (1968). A more general definition of it is $\pi_{ab} = \alpha_a \beta_b$ for some fixed set of cells. See Caussinus (1966), Fienberg (1970b, 1972), and Goodman (1968). Caussinus used the concept to analyze tables that deleted a certain set of cells from consideration, and Goodman used it in earlier analyses of social mobility. Stigler (1999, Chap. 19) summarized early uses, including Karl Pearson's handling in 1913 of a triangular array. Booth and Butler (1999), Krampe et al. (2011), and Smith et al. (1996) presented exact tests for square-table models.

11.4 Occupational mobility: Models for square tables are applied often to the study of occupational (or social) mobility. Each observation pairs parent's occupation with child's occupation. See Goodman (1979b), Hout et al. (1987), Sobel et al. (1998), and Xie (1992).

11.5 Upper-triangular tables: In some applications a table is a priori symmetric or independent, but only the pair (ij) rather than their order is observable, thus leading to an upper-triangular table (Altham 1975). See Khamis (1983) for examples and ML fitting of models for such three-way tables that are symmetric within layers.

Section 11.5: Measuring Agreement Between Observers

11.6 Kappa, intraclass correlation, QS: Banerjee et al. (1999) and Fleiss et al. (2003, Chap. 18) reviewed kappa and its generalizations. See also Kraemer et al. (2002), who discussed versions of kappa and its generalizations that should or should not be used, and Landis et al. (2011), who described a concordance index that has kappa measures as special cases. Kappa and weighted kappa relate to the *intraclass correlation*, a measure of interrater reliability for interval scales (Fleiss 1981; Fleiss and Cohen 1973; Kraemer 1979, Landis et al. 2011). Agresti (2010, Sec. 8.5.3), Becker and Agresti (1992), Goodman (1979b), and Tanner and Young (1985) modeled agreement using loglinear models. Darroch and McCloud (1986) showed that quasi-symmetry has an important role in agreement modeling.

Section 11.6: Bradley–Terry Model for Paired Preferences

11.7 Bradley–Terry generalized: Zermelo (1929) proposed a model that is equivalent to the Bradley-Terry model. Luce (1959) provided an axiomatic basis for it. Mosteller (1951) and Thurstone (1927) proposed an analogous model with the probit link. An interview of Ralph Bradley by M. Hollander (*Statist. Sci.* **16**: 75–100, 2001) discussed food-tasting applications that motivated its development. For extensions, see Bradley (1976), David (1988), and Imrey (2005). Fienberg and Larntz (1976) and Imrey et al. (1976) related it to quasi-independence. Dudbridge (2007) suggested it in genetics for modeling pairs of transmitted/nontransmitted alleles for multiallelic markers. Dittrich et al. (1998) allowed covariates. Matthews and Morris (1995) gave an application with a factorial design, ties, and allowance for dependence among judgments. Böckenholt and Dillon (1997) and Dittrich et al. (2007) modeled dependence with ordinal preferences.

Section 11.7: Marginal Models and Quasi-symmetry Models for Matched Sets

11.8 MH/CMH: Darroch (1981) surveyed thoroughly the relationships among statistics for testing marginal homogeneity and their connections with generalized CMH analyses. See also Mantel and Byar (1978) and White et al. (1982). Bergsma et al. (2009) considered several hypotheses for longitudinal data in the context of the generalized loglinear model [\(10.10\)](#).

EXERCISES

Applications

11.1 A poll by Louis Harris and Associates of 1249 adult Americans indicated that 36% believe in ghosts and 37% believe in astrology. Can you compare these proportions inferentially? If yes, do so. If not, explain what further information you need.

11.2 Refer to [Table 11.1](#) for the 2004 and 2008 Presidential elections. The corresponding 2010 GSS sample for the 585 females had counts 266 and 13 in row 1 and 94 and 212 in row 2. Conduct all steps of McNemar's test, and interpret.

11.3 [Table 11.14](#) shows data about belief in heaven and belief in hell.

Table 11.14 Belief in Heaven and Belief in Hell, for Exercise 11.3

Belief in Heaven	Belief in Hell		
	Yes	No	Total
Yes	955	162	1117
No	9	188	197
Total	964	350	1314

Source: 2008 General Social Survey.

- Compare the marginal proportions using a 95% confidence interval.
- Perform McNemar's test, and interpret results.
- Explain how these data suggest that the marginal proportions are strongly dependent, rather than independent. Explain why inferences in (a) are more precise than if we had the same sample proportions but with independent samples of size 1314 each.

11.4 In the 2006 GSS, subjects were asked their opinion about whether the federal government should fund stem cell research. In 2001, President George W. Bush had instituted a policy that barred the NIH from funding research on embryonic stem cells beyond using the existing cell lines. For those who responded that the government should definitely fund such research, was there a change between 2000 and 2004 in the relative numbers voting for Bush? For the 152 GSS subjects in 2006 who voted Democrat or Republican in each election and who supported funding stem cell research, 89 voted Democrat each time, 52 voted Republican each time, 7 voted Democrat in 2000 and Republican in 2004, and 4 voted Republican in 2000 and Democrat in 2004. Compare the marginal proportions using a 95% confidence interval, and perform the small-sample analog of McNemar's test. Interpret.

11.5 Refer to [Table 9.16](#) and Exercise 9.1. Treat the data as matched pairs on opinion, stratified by gender. Testing independence for the 2×2 table using entries (6, 160) in row 1 and (11, 181) in row 2 tests equality of β for logistic model [\(11.8\)](#) for each gender. Explain why.

11.6 A crossover experiment with 100 subjects compares two drugs for treating migraine headaches. The response scale is success (1) or failure (0). Half the study subjects, randomly selected, used drug *A* the first time they had a headache and drug *B* the next time. For them, 6 responded (1,1) for (*A,B*), 25 responded (1,0), 10 responded (0,1), and 9 responded (0,0). For the 50 subjects who took the drugs in the reverse order, 10 were (1,1) for (*A,B*), 20 were (1,0), 12 were (0,1), and 8 were (0,0).

- Ignoring treatment order, compare the success probabilities for the two drugs. Interpret.
- McNemar's test uses only the pairs of outcomes that differ. For this study, [Table 11.15](#) shows such data from both treatment orders. Testing independence for this table tests whether success rates are identical for the treatments (Gart 1969). Explain why. Analyze these data, and interpret.

Table 11.15 Data for Exercise 11.6

Treatment That Is Better	
Treatment Order First	Second

<i>A</i> , then <i>B</i>	25	10
<i>B</i> , then <i>A</i>	12	10

11.7 A case-control study has 8 pairs of subjects. The cases have colon cancer, and the controls are matched with the cases on gender and age. A possible explanatory variable is the extent of red meat in a subject's diet, measured as "1 = high" or "0 = low." The (case, control) observations on this were (1,1) for 3 pairs, (0,0) for 1 pair, (1,0) for 3 pairs, and (0,1) for 1 pair.

- a. Cross-classify the 8 pairs in terms of diet (1 or 0) for the case against diet (1 or 0) for the control. Call this Table A. Display the $2 \times 2 \times 8$ table with eight partial tables relating diet (1 or 0) to response (case or control) for the 8 pairs. Call this Table B.
- b. Calculate the McNemar z^2 for Table A and the CMH statistic for Table B. Compare.
- c. Show that the Mantel-Haenszel estimate (6.7) of a common odds ratio for Table B is identical to n_{12}/n_{21} for Table A.
- d. For Table B with pairs deleted in which the case and the control had the same diet, show that the CMH statistic and the Mantel-Haenszel odds ratio estimate do not change.
- e. This sample size is small for large-sample tests. Use the binomial distribution with Table A to find the exact two-sided P -value.

11.8 For Table 11.14 above, find the estimated odds ratio for comparing the distributions for belief in heaven and for belief in hell, using (a) model (11.6) with a population-averaged effect and (b) model (11.8) with a subject-specific effect. (c) Explain why they differ.

11.9 Table 11.16 shows subjects' purchase choice of instant decaffeinated coffee at two times.

Table 11.16 Data for Exercise 11.9 on Coffee Purchases

First Purchase	Second Purchase				
	High Point	Taster's Choice	Sanka	Nescafe	Brim
High Point	93	17	44	7	10
Taster's Choice	9	46	11	0	9
Sanka	17	11	155	9	12
Nescafe	6	4	9	15	2
Brim	10	4	12	2	27

Source: Based on data from R. Grover and V. Srinivasan, *J. Market. Res.* 24: 139–153, 1987. Reprinted with permission from the American Marketing Association.

- a. Fit the symmetry model and use residuals to analyze changes.
- b. Test marginal homogeneity. Show that the small P -value reflects a decrease in the proportion choosing High Point and an increase in the proportion choosing Sanka, with no evidence of change for the other coffees.
- c. Show that quasi-independence has $G^2 = 13.8$ ($df = 11$). Interpret, and suggest other analyses that might be useful.

11.10 For Table 11.6, fit the ordinal quasi-symmetry model using unequally spaced but sensible scores. Compare results and interpretations to those in Sections 11.3.4 and 11.4.7.

11.11 Table 11.17 relates father's and son's occupational status for a British sample. Analyze these data, using models of (a) symmetry, (b) quasi-symmetry, (c) ordinal quasi-symmetry, (d) marginal homogeneity, and (e) quasi-independence. Interpret their fit and lack of fit.

Table 11.17 Occupational Mobility Data for Exercise 11.11

Father's Status	Son's Status				
	1	2	3	4	5
1	50	45	8	18	8
2	28	174	84	154	55
3	11	78	110	223	96
4	14	150	185	714	447
5	3	42	72	320	411

Source: Reprinted with permission from D. V. Glass (ed.), *Social Mobility in Britain*, Glencoe, IL: Free Press, 1954.

11.12 Each week *Variety* magazine summarizes reviews of new movies by critics in several cities. Each review is categorized as pro, con, or mixed, according to whether the overall evaluation is positive, negative, or a mixture of the two. [Table 11.18](#) summarizes the ratings from April 1995 through September 1996 for Chicago film critics Gene Siskel and Roger Ebert.

[Table 11.18](#) Data for Exercise 11.12 on Movie Reviews

		Ebert		
		Con	Mixed	Pro
Siskel				
Con	24	8	13	
Mixed	8	13	11	
Pro	10	9	64	

Source: A. Agresti and L. Winner, *CHANCE* 10: 10–14, 1997. Reprinted with permission, copyright 1997 by the American Statistical Association.

- a. Fit the symmetry model, quasi-independence model, and quasi-symmetry model. Interpret.
- b. Test marginal homogeneity using models, and interpret.
- c. Summarize these data using the kappa measure of agreement.

11.13 [Table 11.19](#) displays multiple sclerosis diagnoses for two neurologists who classified patients in Winnipeg and in New Orleans with the scale (1) certain, (2) probable, (3) possible, and (4) doubtful, unlikely, or definitely not. For the New Orleans patients, study the agreement using (a) the independence model and residuals, (b) more complex models, and (c) kappa. Interpret each.

[Table 11.19](#) Data for Exercise 11.13 on Neurologist Agreement

New Orleans Neurologist	Winnipeg Neurologist							
	Winnipeg Patients				New Orleans Patients			
	1	2	3	4	1	2	3	4
1	38	5	0	1	5	3	0	0
2	33	11	3	0	3	11	4	0
3	10	14	5	6	2	13	3	4
4	3	7	3	10	1	2	4	14

Source: J. R. Landis and G. G. Koch, *Biometrics* 33: 159–174, 1977. Reprinted with permission from the Biometric Society.

11.14 For Exercise 11.13, construct a model that describes agreement between neurologists for the two sites simultaneously.

11.15 Calculate kappa for the 4×4 table having $n_{ii} = 5$ all i , $n_{i,i+1} = 15$, $i = 1, 2, 3$, $n_{41} = 15$, and $n_{ij} = 0$ otherwise. Use these data to explain why strong association does not imply strong agreement.

11.16 In 1990, a sample of psychology graduate students at the University of Florida made blind, pairwise preference tests of three cola drinks. For 49 comparisons of Coke and Pepsi, Coke was preferred 29 times. For 47 comparisons of Classic Coke and Pepsi, Classic Coke was preferred 19 times. For 50 comparisons of Coke and Classic Coke, Coke was preferred 31 times. Comparisons resulting in ties are not reported.

- a. Fit the Bradley–Terry model, analyze the quality of fit, and rank the drinks. Is there sufficient evidence to conclude a preference for one drink?
- b. Estimate the probability that Coke is preferred to Pepsi, using the model, and compare to the sample proportion.

11.17 [Table 11.20](#) refers to journal citations among four statistics journals during 1987–1989. The more often articles in a particular journal are cited, the more prestige that journal accrues. For citations involving pair A and B , view it as a victory for A if it is cited by B and a defeat for A if it cites B . Fit the Bradley–Terry model. Interpret the fit, and give a prestige ranking of the journals. For citations involving *Communications in Statistics* and *JRSS-B*, estimate the probability that the *Commun. Stat.* article cites the *JRSS-B* article.

[Table 11.20](#) Data for Exercise 11.17 on Journal Citations

Citing Journal	Cited Journal			
	Biometrika	Commun. Stat.	JASA	JRSS-B
Biometrika	714	33	320	284
Commun. Stat.	730	425	813	276
JASA	498	68	1072	325
JRSS-B	221	17	142	188

Source: Stigler (1994). Reprinted with permission from the Institute of Mathematical Statistics.

11.18 [Table 11.21](#) refers to matches for several men tennis players between January 2009 and June 2011.

Table 11.21 Data for Exercise 11.18 on 2009–2011 Men’s Tennis Results

Winner	Loser				
	Nadal	Federer	Murray	Roddick	Djokovic
Nadal	—	5	6	2	6
Federer	2	—	4	5	7
Murray	3	4	—	2	1
Roddick	1	0	1	—	4
Djokovic	7	7	2	1	—

Source: www.atptowntour.com.

a. Fit the Bradley–Terry model. Interpret, and rank the players. Explain why it is probably not realistic, however, to treat the matches between a particular pair of players as all having the same probabilities. [Hint: They were played on quite varied surfaces, including clay and grass.]

b. Estimate the probability of Nadal beating Federer. Compare the model estimate to the sample proportion. Construct a 90% confidence interval for the probability.

11.19 Refer to Exercise 3.7 on basketball free-throw shooting. Analyze these data.

11.20 Refer to Exercise 9.5. The two-way table relating responses for government spending on the environment (as rows) and cities (as columns) has cell counts, by row, (108, 179, 157 / 21, 55, 52 / 5, 6, 24). Analyze these data.

11.21 Analyze the data in [Table 2.13](#) on sexual attitudes with methods presented in this chapter.

11.22 [Table 11.22](#) comes from a crossover study in which each subject used each of three drugs for treatment of a chronic condition at three times. Show that the sample proportion favorable was (0.61, 0.61, 0.35) for drugs (A, B, C), and that the likelihood-ratio statistic for testing marginal homogeneity is 5.95 ($df = 2$), for a P -value of 0.051. Use the Bonferroni method to find simultaneous 95% Wald confidence intervals for the differences between proportions for pairs of treatments.

Table 11.22 Crossover Study Results for Exercise 11.22

	Drug A Favorable		Drug A Unfavorable	
	B Favorable	B Unfavorable	B Favorable	B Unfavorable
C Favorable	6	2	2	6
C Unfavorable	16	4	4	6

Source: Reprinted with permission from the Biometric Society (Grizzle et al. 1969).

[Applying Bonferroni with the score confidence interval gives¹ $(-0.015, 0.496)$ for drugs A and C and for drugs B and C, and $(-0.195, 0.195)$ for drugs A and B.]

11.23 Refer to [Table 9.3](#). Viewing the table as matched triplets, construct the marginal distribution for each of marijuana, alcohol, and cigarettes. Test the hypothesis of marginal homogeneity. Interpret results.

11.24 Refer to [Table 11.1](#) and Exercise 11.2. Regarding the data for both males and females as a $2 \times 2 \times 2$ table, use loglinear models to describe the associations between gender and Presidential vote in each year.

Theory and Methods

11.25 In genetics, the transmission/disequilibrium test (TDT) considers the transmission of a variant allele of a biallelic marker from heterozygous parents to affected children (Dudbridge 2007). It treats the untransmitted allele as a matched control to the transmitted allele. Show how to express the hypothesis that heterozygous parents transmit the two alleles with equal probability in the context of a table with row and category columns (variant, common). Explain why McNemar's test is appropriate for testing that hypothesis.

11.26 Explain the following analogy: McNemar's test is to binary data as the paired difference t test is to normally distributed data.

11.27 For a 2×2 table, derive $\text{cov}(p_{+1}, p_{1+})$, and show that $\text{var}[\sqrt{n}(p_{+1} - p_{1+})]$ equals (11.1).

11.28 Consider the subject-specific model (11.8) for binary matched pairs.

- a. Show that $\exp(\beta)$ is a conditional odds ratio between observation and outcome. Explain the distinction between it and the odds ratio $\exp(\beta)$ for model (11.6).
- b. Using the conditional distribution (11.9), show that $\hat{\beta} = \log(n_{21}/n_{12})$.
- c. Use the delta method to show (11.10) for the SE of $\hat{\beta}$.
- d. Averaging over the population, explain why

$$\pi_{21} = E\left[\frac{1}{1 + \exp(\alpha_i)} \frac{\exp(\alpha_i + \beta)}{1 + \exp(\alpha_i + \beta)}\right],$$

where the expectation is with respect to the distribution for $\{\alpha_i\}$. Similarly, state π_{12} . For a random sample of size n , explain why as $n \rightarrow \infty$, $n_{21}/n_{12} = p_{21}/p_{12} \xrightarrow{P} \exp(\beta)$. [Hint: Apply the law of large numbers due to A. A. Markov for independent but not identically distributed random variables, or use Chebyshev's inequality.]

- e. Show that the Mantel–Haenszel estimator (6.7) of a common odds ratio in the $2 \times 2 \times n$ subject-specific table simplifies to $\exp(\hat{\beta}) = n_{21}/n_{12}$ for the population-averaged table.
- f. Show that the CMH statistic (6.6) for a $2 \times 2 \times n$ subject-specific table is algebraically identical to the McNemar statistic $(n_{21} - n_{12})^2/(n_{21} + n_{12})$ for the population-averaged table.

11.29 Refer to Exercise 11.28. Unlike the conditional ML estimator of β , the unconditional ML estimator is inconsistent (Andersen 1980, pp. 244–245; first shown by him in 1973). Show this as follows:

- a. Assuming independence of responses for different subjects and different observations by the same subject, find the log likelihood. Show that the likelihood equations are $y_{+t} = \sum_i P(Y_{it} = 1)$ and $y_{i+} = \sum_t P(Y_{it} = 1)$.
- b. Substituting $\exp(\alpha_i)/[1 + \exp(\alpha_i)] + \exp(\alpha_i + \beta)/[1 + \exp(\alpha_i + \beta)]$ in the second likelihood equation, show that $\hat{\alpha}_i = -\infty$ for the n_{22} subjects with $y_{i+} = 0$, $\hat{\alpha}_i = \infty$ for the n_{11} subjects with $y_{i+} = 2$, and $\hat{\alpha}_i = -\hat{\beta}/2$ for the $n_{21} + n_{12}$ subjects with $y_{i+} = 1$.
- c. By breaking $\sum_i P(Y_{it} = 1)$ into components for the sets of subjects having $y_{i+} = 0$, $y_{i+} = 2$, and $y_{i+} = 1$, show that the first likelihood equation is, for $t = 1$, $y_{+1} = n_{22}(0) + n_{11}(1) + (n_{21} + n_{12})\exp(-\hat{\beta}/2)/[1 + \exp(-\hat{\beta}/2)]$. Explain why $y_{+1} = n_{11} + n_{12}$, and solve the first likelihood equation to show that $\hat{\beta} = 2\log(n_{21}/n_{12})$. Hence, as a result of Exercise 11.28, $\hat{\beta} \xrightarrow{P} 2\beta$.

11.30 Consider marginal model (11.6) when Y_1 and Y_2 are independent and subject-specific model (11.8) when $\{\alpha_i\}$ are identical. Explain why they are equivalent.

11.31 Let $\hat{\beta}_M = \log(p_{+1} p_{2+}/p_{+2} p_{1+})$ refer to marginal model (11.6) and $\hat{\beta}_C = \log(n_{21}/n_{12})$ to subject-specific model (11.8). Using the delta method, show that the asymptotic variance of $\sqrt{n}(\hat{\beta}_M - \beta_M)$ is

$$(\pi_{1+} \pi_{2+})^{-1} + (\pi_{+1} \pi_{+2})^{-1} - 2(\pi_{11} \pi_{22} - \pi_{12} \pi_{21})/(\pi_{1+} \pi_{2+} \pi_{+1} \pi_{+2}).$$

Under the independence condition of the previous exercise, $\beta_M = \beta_C$. In that case, show that the asymptotic variances satisfy

$$\begin{aligned}\text{var}[\sqrt{n}(\hat{\beta}_M)] &= (\pi_{1+}\pi_{2+})^{-1} + (\pi_{+1}\pi_{+2})^{-1} \\ &\leq (\pi_{1+}\pi_{+2})^{-1} + (\pi_{+1}\pi_{2+})^{-1} = \pi_{12}^{-1} + \pi_{21}^{-1} = \text{var}[\sqrt{n}(\hat{\beta}_C)].\end{aligned}$$

11.32 For model (11.12) for a matched-pairs study, with the conditional ML approach show that the conditional distribution satisfies (11.13) and does not depend on β when $S_i = 0$ or 2. Show what happens to β_j in the conditional distribution for a predictor for which $x_{ji1} = x_{ji2}$ for all i .

11.33 Consider model (11.12) for a study with matched sets of T observations rather than matched pairs. Explain how (11.13) generalizes and construct the form of the conditional likelihood.

11.34 Give an example illustrating that when $I > 2$, marginal homogeneity does not imply symmetry.

11.35 Derive the likelihood equations and residual df for (a) symmetry, (b) quasi-symmetry, and (c) quasi-independence.

11.36 For the quasi-symmetry model (11.19), let $\lambda_a = \lambda_a^x - \lambda_a^r$. Show that the model can be expressed equivalently as $\log \mu_{ab} = \lambda + \lambda_a + \lambda_{ab}^*$, with $\lambda_{ab}^* = \lambda_{ba}^*$. Hence, we need only one set of main-effect parameters.

11.37 Show that quasi-symmetry is equivalent (Caussinus 1966) to

$$(\pi_{ab}\pi_{bc}\pi_{ca})/(\pi_{ba}\pi_{cb}\pi_{ac}) = 1 \quad \text{all } a, b, \text{ and } c.$$

11.38 Derive the covariance matrix V for the difference vector d that is estimated in expression (11.15).

11.39 Construct the loglinear model satisfying both marginal homogeneity and statistical independence. Show that $\hat{\pi}_{ab} = (p_{+a} + p_{a+})(p_{+b} + p_{b+})/4$ and residual df = $I(I - 1)$.

11.40 Identify loglinear models that correspond to the logistic models, for $a < b$, $\log(\pi_{ab}/\pi_{ba}) =$ (a) 0, (b) τ , (c) $\alpha_a - \alpha_b$, and (d) $\beta(b - a)$.

11.41 Consider the multiplicative model for a square table,

$$\pi_{ab} = \begin{cases} \alpha_a \alpha_b (1 - \beta), & a \neq b \\ \alpha_a^2 + \beta \alpha_a (1 - \alpha_a), & a = b. \end{cases}$$

a. Show that the model satisfies (i) symmetry, (ii) marginal homogeneity, (iii) quasi-symmetry, and (iv) quasi-independence.

b. Show that $\alpha_a = \pi_{a+} = \pi_{+a}$, $a = 1, \dots, I$.

c. Show that $\beta = \text{Cohen's kappa}$, and interpret $\beta = 0$ and $\beta = 1$ for this model.

11.42 A 2×2 table has a true odds ratio of 10. Find the cell probabilities for which (a) $\pi_{1+} = \pi_{+1} = 0.50$ and (b) $\pi_{1+} = \pi_{+1} = 0.10$. Find kappa for each. (This shows that for a given association, kappa depends strongly on the marginal probabilities.)

11.43 Consider the Bradley-Terry model (11.30).

a. Show that $\log(\Pi_{ac}/\Pi_{ca}) = \log(\Pi_{ab}/\Pi_{ba}) + \log(\Pi_{bc}/\Pi_{cb})$.

b. With this model, is it possible that a could be preferred to b (i.e., $\Pi_{ab} > \Pi_{ba}$) and b could be preferred to c , yet c could be preferred to a ? Explain.

c. Explain why $\{\beta_a\}$ are not identifiable without a constraint such as $\beta_I = 0$. [Hint: Show the model holds when $\{\beta_a^*\} = \{\beta_a - c\}$ for any c .]

11.44 Refer to model (11.31) for baseball home-team advantage.

a. Construct a more general model having home-team parameters $\{\beta_{Hi}\}$ and away-team parameters $\{\beta_{Ai}\}$, such that the probability team i beats team j when i is the home team is $\exp(\beta_{Hi})/[\exp(\beta_{Hi}) + \exp(\beta_{Aj})]$. where $\beta_{Ai} = 0$ but β_{Hi} is unrestricted.

b. Interpret the case $\{\beta_{Hi} = \beta_{Ai} + c\}$, when (i) $c = 0$ and (ii) $c > 0$.

c. Fit the model to Table 11.12. Compare the fit to model (11.31). Compare $\{\hat{\beta}_{Hi}\}$ and $\{\hat{\beta}_{Ai}\}$ to describe how teams play at home and away.

11.45 Find the log likelihood for the Bradley-Terry model. From the kernel, show that (given

$\{N_{ab}\})$ the minimal sufficient statistics are $\{n_{a+}\}$. Thus, explain how “victory totals” determine the estimated ranking.

11.46 Explain how to fit the complete symmetry model in T dimensions.

11.47 Prove that if k th-order marginal symmetry holds, then j th-order marginal symmetry holds for any $j < k$.

11.48 Suppose quasi-symmetry holds for an I^T table. When the table is collapsed over a variable, show that the model holds for the I^{T-1} table with the same main effects.

11.49 Let $y_{it} = 1$ or 0 for observation t on subject i , $i = 1, \dots, n$, $t = 1, \dots, T$. Let $\bar{y}_{it} = \sum_i y_{it}/n$, $\bar{y}_{i\cdot} = \sum_t y_{it}/T$, and $\bar{y} = \sum_i \sum_t y_{it}/nT$. Regard $\{y_{i+}\}$ as fixed, and suppose each way to allocate the y_{i+} “successes” to y_{i+} of the observations is equally likely.

a. Show that $E(Y_{it}) = \bar{y}_{i\cdot}$, $\text{var}(Y_{it}) = \bar{y}_{i\cdot}(1 - \bar{y}_{i\cdot})$, and $\text{cov}(Y_{it}, Y_{ik}) = -\bar{y}_{i\cdot}(1 - \bar{y}_{i\cdot})/(T - 1)$ for $t \neq k$.

[Hint: The covariance is the same for any pair of cells in the same row, and $\text{var}(\sum_t Y_{it}) = 0$ since y_{i+} is fixed.]

b. For large n with independent subjects, explain why $(\bar{Y}_{\cdot 1}, \dots, \bar{Y}_{\cdot T})$ is approximately multivariate normal with pairwise correlation $\rho = -1/(T - 1)$. Conclude that Cochran’s Q statistic (Cochran 1950)

$$Q = \frac{n^2(T - 1) \sum_{i=1}^T (\bar{y}_{i\cdot} - \bar{y})^2}{T \sum_{i=1}^n \bar{y}_{i\cdot}(1 - \bar{y}_{i\cdot})}$$

is approximately chi-squared with $\text{df} = (T - 1)$. [One way notes that if (X_1, \dots, X_T) is multivariate normal with common mean and common variance σ^2 and common correlation ρ for pairs (X_p, X_k) , then $\sum_t (X_t - \bar{X})^2 / \sigma^2(1 - \rho)$ is chi-squared with $\text{df} = (T - 1)$. Bhapkar and Somes (1977) gave slightly weaker conditions for a chi-squared limiting distribution for Q .]

c. Show that Q is unaffected by deleting cases in which $y_{i1} = \dots = y_{iT}$.

¹Thanks to Bernhard Klingenberg for these results.

CHAPTER 12

Clustered Categorical Data: Marginal and Transitional Models

Many studies observe the response variable for each subject repeatedly, at several times or under various conditions. Repeated categorical response data occur commonly in health-related applications, especially in longitudinal studies. For example, a physician might evaluate patients taking a new drug or a placebo at several occasions regarding whether the treatment is successful.

In the next three chapters we present models that apply to data in which repeated observations occur for matched sets, or *clusters*, of observations. In a longitudinal study, a cluster consists of the set of repeated observations over time by a particular subject. But the clustered responses need not refer to different times. A dental study might measure whether there is decay for each tooth in a subject's mouth. The set of teeth within a subject's mouth form a cluster. A study of factors that affect children's weight, measured as (normal, overweight, obese), might sample families and treat children from the same family as a cluster. A toxicity study may observe a (survival, nonsurvival) response for each fetus in a litter, for a sample of pregnant mice exposed to various dosages of a toxin. Each litter forms a cluster.

In such applications, observations within a cluster tend to be more alike than observations from different clusters. Ordinary analyses that ignore the correlation usually have invalid standard errors. Aitkin et al. (1981) gave an example from a project about teaching styles where this makes a substantive difference, the statistical significance of certain differences being substantially reduced after allowing for correlation among children taught by the same teacher. In Section 11.1.4 we noted that positive correlation between sample proportions results in improved precision for estimating within-subject effects. By contrast, for inference about between-subject effects (such as comparing genders or races), T repeated observations for a single subject do not provide as much information as T observations on different subjects, and positive correlations result in larger standard errors for such effects (Exercise 12.21).

In this chapter we generalize the marginal model methods of Chapter 11 for matched pairs to clustered data also having explanatory variables, such as a study that compares the distribution of repeated observations for different groups or treatments. In Section 12.1 we introduce marginal models with explanatory variables and fit them using ML methods. In Section 12.2 we fit marginal models by solving *generalized estimating equations* (GEEs). This method is a multivariate version of quasi-likelihood that is computationally simpler than ML. Section 12.3 presents technical details about the GEEs approach. In Section 12.4 we introduce a *transitional* approach that models each observation in terms of outcomes of other observations, such as *time series* models that use past observations to predict future ones.

12.1 MARGINAL MODELING: MAXIMUM LIKELIHOOD APPROACH

Repeated measurement provides a multivariate response (Y_1, Y_2, \dots, Y_T). Moreover, T often varies by cluster, such as when each cluster is a family or when some observations in a cluster are missing. In this section we consider marginal models for the $\{Y_t\}$, fitted by ML. We defer model fitting details to the end of the section.

12.1.1 Example: Longitudinal Study of Mental Depression

[Table 12.1](#) refers to a longitudinal study comparing a new drug with a standard drug for treatment of 340 subjects suffering mental depression (Koch et al. 1977). Subjects were classified into two initial diagnosis groups according to whether severity of depression was mild or severe. In each group, subjects were randomly assigned to one of the two drugs. Following 1 week, 2 weeks, and 4 weeks of treatment, each subject's suffering from mental depression was classified as normal or abnormal.

Table 12.1 Cross-Classification of Responses on Depression at Three Times, by Diagnosis and Treatment

Diagnosis	Treatment	Response at Three Times ^a							
		NNN	NNA	NAN	NAA	ANN	ANA	AAN	AAA
Mild	Standard	16	13	9	3	14	4	15	6
	New drug	31	0	6	0	22	2	9	0
Severe	Standard	2	2	8	9	9	15	27	28
	New drug	7	2	5	2	31	5	32	6

^aN, normal; A, abnormal.

Source: Reprinted with permission from the Biometric Society (Koch et al. 1977).

[Table 12.1](#) shows four groups, the combinations of categories of the two explanatory variables: treatment type and severity of initial diagnosis. Since the study observed the binary response (depression assessment) at $T = 3$ occasions, [Table 12.1](#) contains cell counts for a 23 contingency table for each group. The three depression assessments form a multivariate response variable with three components. The 12 marginal distributions result from three repeated observations for each of the four groups.

Let t denote the time of measurement. Denote observation t for a subject by $Y_t = 1$ for normal and $Y_t = 0$ for abnormal. Let s denote the severity of the initial diagnosis, with $s = 1$ for severe and $s = 0$ for mild. Let d denote the drug, with $d = 1$ for new and $d = 0$ for standard. Koch et al. (1977) noted that if the time metric reflects cumulative drug dosage, a logit scale often has a linear effect for the logarithm of time. They used scores (0, 1, 2) for t , the logs to base 2 of the week numbers (1, 2, and 4).

[Table 12.2](#) shows sample proportions of normal responses for the 12 marginal distributions. For instance, from [Table 12.1](#), the sample proportion of normal responses after week 1 for subjects with mild initial severity using the standard drug was

Table 12.2 Sample Marginal Proportions of Normal Response for Depression Data of [Table 12.1](#)

Diagnosis	Treatment	Sample Proportion		
		Week 1	Week 2	Week 4
Mild	Standard	0.51	0.59	0.68
	New drug	0.53	0.79	0.97
Severe	Standard	0.21	0.28	0.46
	New drug	0.18	0.50	0.83

$$(16 + 13 + 9 + 3)/(16 + 13 + 9 + 3 + 14 + 4 + 15 + 6) = 0.51.$$

The sample proportion of normal responses (1) increased over time for each group; (2) increased at a faster rate for the new drug than the standard, for each fixed severity; and (3) was higher for the mild than the severe initial severity diagnosis, for each treatment at each occasion. In such a study the company that developed the new drug would hope to show that patients have a significantly higher rate of improvement with that drug.

The marginal logistic model

$$\text{logit}[P(Y_t = 1)] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t$$

has main effects for the explanatory variables (severity and drug) and the variable (time) that specifies the different components of the multivariate response. Its linear time effect β_3 is the same

for each group.

The natural sampling assumption is multinomial for the eight cells in the 2^3 cross-classification of the three responses, independently for the four groups. However, the model refers to 12 marginal probabilities (for 2 drug treatments \times 2 severity diagnoses \times 3 time points) rather than the $4 \times 23 = 32$ cell probabilities in the product multinomial likelihood function. The three marginal binomial variates for each group are dependent. ML estimation requires an iterative routine for maximizing the product multinomial likelihood, subject to the constraint that the marginal probabilities satisfy the model. An algorithm for this is described in Section 12.1.4.

A check of model fit compares the 32 cell counts in [Table 12.1](#) to their ML fitted values. Since the model describes 12 marginal logits using four parameters, residual df = 8. The deviance $G^2 = 34.6$. The poor fit is not surprising. The model assumes a common rate of improvement β_3 over time, but [Table 12.2](#) suggests a faster rate for the new drug.

A more realistic model permits the time effect to differ by drug,

$$\text{logit}[P(Y_t = 1)] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4(d \times t).$$

Its ML time effect estimates are $\hat{\beta}_3 = 0.48$ ($SE = 0.12$) for the standard drug ($d = 0$) and $\hat{\beta}_3$ and $\hat{\beta}_4 + 0a = 1.49$ ($SE = 0.14$) for the new one ($d = 1$). For the new drug, the slope is $\hat{\beta}_4 = 1.01$ ($SE = 0.18$) higher than for the standard, giving strong evidence of faster improvement. This model fits much better, with $G^2 = 4.2$ (df = 7). The G^2 decrease of $34.6 - 4.2 = 30.4$ compared to the simpler model is the likelihood-ratio test of $H_0: \beta_4 = 0$, a common time effect for each drug.

The estimate of the severity effect is $\hat{\beta}_1 = -1.29$ ($SE = 0.14$). For each drug \times time combination, the estimated odds of a normal response when the initial diagnosis was severe equal $\exp(-1.29) = 0.27$ times the estimated odds when the initial diagnosis was mild. The estimate $\hat{\beta}_2 = -0.06$ ($SE = 0.22$) indicates an insignificant difference between the drugs after one week (for which $t = 0$). At time t , the estimated odds of normal response with the new drug are $\exp(-0.06 + 1.01 t)$ times the estimated odds for the standard drug, for each severity level. In summary, severity, drug treatment, and time all have substantial effects on the probability of a normal response.

12.1.2 Modeling a Repeated Multinomial Response

Models for marginal distributions of a repeated binary response generalize to multicategory ($I > 2$) responses. At observation t , the marginal response distribution has $(I - 1)$ logits. For a particular marginal logit, a model has the form

$$\text{logit}_j(t) = \alpha_j + \boldsymbol{\beta}_j^T \mathbf{x}_t, \quad j = 1, \dots, I - 1, \quad t = 1, \dots, T.$$

For a nominal response, we can use a baseline-category logit, $\text{logit}_j(t) = \log[P(Y_t = j)/P(Y_t = I)]$, to describe the odds of each outcome relative to a baseline. For ordinal responses, we can use the cumulative logit, $\text{logit}_j(t) = \text{logit}[P(Y_t \leq j)]$. With $\boldsymbol{\beta}_j = \boldsymbol{\beta}$ for all j , the model takes the proportional odds form with the same effects for each logit.

12.1.3 Example: Insomnia Clinical Trial

[Table 12.3](#) shows results of a randomized, double-blind clinical trial comparing an active hypnotic drug with a placebo in 239 patients who have insomnia problems. The response is the patient's reported time to fall asleep after going to bed, grouped into intervals of minutes. Patients responded before and following a two-week treatment period. The two treatments, active and placebo, form a binary explanatory variable. The subjects receiving the two treatments were independent samples.

[Table 12.3](#) Time to Falling Asleep, by Treatment and Occasion

Treatment	Initial	Time to Falling Asleep			
		Follow-up			
		<20	20–30	30–60	>60
Active	<20	7	4	1	0
	20–30	11	5	2	2
	30–60	13	23	3	1
	>60	9	17	13	8
Placebo	<20	7	4	2	1
	20–30	14	5	1	0
	30–60	6	9	18	2
	>60	4	11	14	22

Source: From S. F. Francom, C. Chuang-Stein, and J. R. Landis, *Statist. Med.* **8**: 571–582, 1989. Reprinted with permission from John Wiley & Sons Ltd.

[Table 12.4](#) displays sample marginal distributions for the four treatment \times occasion combinations. From the initial to follow-up occasion, time to falling asleep seems to shift downward for both treatments. The degree of shift seems greater for the active treatment, indicating possible interaction. The response is a discrete version of a continuous variable, so by the derivation in Section 8.2.3 a cumulative link model is natural. The proportional odds version of the cumulative logit model,

$$(12.1) \text{logit}[P(Y_t \leq j)] = \alpha_j + \beta_1 t + \beta_2 x + \beta_3(t \times x),$$

permits interaction between t = occasion (0 = initial, 1 = follow-up) and x = treatment (0 = placebo, 1 = active), but assumes the same effects for each response cutpoint.

For ML model fitting, $G^2 = 8.0$ (df = 6) for comparing observed to fitted cell counts in modeling the 12 marginal logits using these six parameters. The ML estimates are $\hat{\beta}_1 = 1.074$ ($SE = 0.162$), $\hat{\beta}_2 = 0.046$ ($SE = 0.236$), and $\hat{\beta}_3 = 0.662$ ($SE = 0.244$). This shows evidence of interaction. At the initial observation, the estimated odds that time to falling asleep for the active treatment is below any fixed level equal $\exp(0.046) = 1.04$ times the estimated odds for the placebo treatment; at the follow-up observation, the effect is $\exp(0.046 + 0.662) = 2.03$. In other words, initially the two groups had similar distributions, as expected by the randomization of subjects to treatment groups, but at the follow-up those with the active treatment tended to fall asleep more quickly.

For simpler interpretation, it can be helpful to report sample marginal means and their differences. With response scores {10, 25, 45, 75} for time to fall asleep, the initial means were 50.0 for the active group and 50.3 for the placebo. The difference in means between the initial and follow-up responses was 22.2 for the active group and 13.0 for the placebo. The difference between these differences of means equals 9.2, with $SE = 3.0$, indicating that the change was significantly greater for the active group.

12.1.4 ML Fitting of Marginal Logistic Models: Constraints on Cell Probabilities

ML fitting of marginal logistic models is awkward. For T observations on an I -category response, at each setting of predictors the likelihood refers to I^T multinomial joint probabilities, but the model applies to T sets of marginal multinomial parameters $\{P(Y_t = k), k = 1, \dots, I\}$.

Table 12.4 Sample Marginal Distributions for Insomnia Data of [Table 12.3](#)

Treatment	Occasion	Time to Falling Asleep			
		<20	20–30	30–60	>60
Active	Initial	0.101	0.168	0.336	0.395
	Follow-up	0.336	0.412	0.160	0.092
Placebo	Initial	0.117	0.167	0.292	0.425
	Follow-up	0.258	0.242	0.292	0.208

Let $\boldsymbol{\pi}$ denote the complete set of multinomial joint probabilities for all settings of predictors. Marginal logistic models have the generalized loglinear model form

$$(12.2) \quad \mathbf{C} \log(\mathbf{A}\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}$$

introduced in Section 10.5.1. In the binary case, the matrix \mathbf{A} applied to $\boldsymbol{\pi}$ forms the T marginal probabilities $\{P(Y_t = 1)\}$ and their complements at each setting of predictors. The matrix \mathbf{C} applied to the log marginal probabilities forms the T marginal logits for each setting; each row of \mathbf{C} has 1 in the position multiplied by the log numerator probability for a given marginal logit, -1 in the position multiplied by the log denominator probability, and 0 elsewhere.

For instance, for the model of marginal homogeneity in a 2^T table with no covariates,

$$\text{logit}[P(Y_t = 1)] = \alpha, \quad t = 1, \dots, T,$$

$\boldsymbol{\beta}$ is a single parameter, denoted by α here. For $T = 2$, $\boldsymbol{\pi}$ has four elements, and this model is

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \log \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{21} \\ \pi_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \alpha,$$

setting both $\text{logit}(\pi_{11} + \pi_{12}) = \text{logit}[P(Y_1 = 1)]$ and $\text{logit}(\pi_{11} + \pi_{21}) = \text{logit}[P(Y_2 = 1)]$ equal to α .

The likelihood function $\ell(\boldsymbol{\pi})$ for a marginal logistic model is the product of the multinomial mass functions from the various predictor settings. One approach for ML fitting views the model as a set of constraints and uses methods for maximizing a function subject to constraints. In model (12.2), let \mathbf{U} denote a full column rank matrix such that the space spanned by the columns of \mathbf{U} is the orthogonal complement of the space spanned by the columns of \mathbf{X} . Then, $\mathbf{U}^T \mathbf{X} = \mathbf{0}$, and the model has the equivalent constraint form

$$\mathbf{U}^T \mathbf{C} \log(\mathbf{A}\boldsymbol{\pi}) = \mathbf{0}.$$

For instance, for marginal homogeneity in a 2×2 table with (12.2) as expressed above, $\mathbf{U}^T = (1, -1)$. Then, \mathbf{U}^T applied to $\mathbf{C} \log(\mathbf{A}\boldsymbol{\pi})$ sets the difference between the row and column marginal logits equal to 0.

This method of maximizing the likelihood incorporates these model constraints as well as identifiability constraints, which constrain the response probabilities at each predictor setting to sum to 1. We express this collection of model constraints $\mathbf{U}^T \mathbf{C} \log(\mathbf{A}\boldsymbol{\pi}) = \mathbf{0}$ and identifiability constraints as $\mathbf{f}(\boldsymbol{\pi}) = \mathbf{0}$. The method introduces Lagrange multipliers corresponding to these constraints and solves the Lagrangian likelihood equations using a Newton–Raphson algorithm (Aitchison and Silvey 1958, Haber 1985). Let $\boldsymbol{\theta}$ be a vector having elements $\boldsymbol{\pi}$ and the Lagrange multipliers $\boldsymbol{\lambda}$. The Lagrangian likelihood equations have form $\mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$, where

$$\mathbf{h}(\boldsymbol{\theta}) = \mathbf{h}(\boldsymbol{\pi}, \boldsymbol{\lambda}) = (\mathbf{f}(\boldsymbol{\pi}), \partial \log[\ell(\boldsymbol{\pi})]/\partial \boldsymbol{\pi} + [\partial \mathbf{f}(\boldsymbol{\pi})/\partial \boldsymbol{\pi}]^T \boldsymbol{\lambda})^T$$

is a vector with terms involving the contrasts in marginal logits that the model specifies as constraints as well as log-likelihood derivatives.

The Newton–Raphson method then applies as

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \left[\frac{\partial h(\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\theta}} \right]^{-1} h(\boldsymbol{\theta}^{(t)}), \quad t = 1, \dots$$

This can be computationally intensive because the derivative matrix inverted has dimensions larger than the number of elements in $\boldsymbol{\pi}$. A refinement (Lang 1996a, Lang and Agresti 1994) uses an asymptotic approximation to a reparameterized derivative matrix that has a much simpler form, requiring inverting only a diagonal matrix and a symmetric positive definite matrix.

This ML marginal fitting method is available in specialized software. (The computing Appendix at the text website describes an R function, *mph.fit*, developed by J. Lang.) The method makes no assumption about the model that describes the joint distribution $\boldsymbol{\pi}$. Thus, when the marginal model holds, the ML estimate of $\boldsymbol{\beta}$ in (12.2) is consistent regardless of the dependence structure for that distribution.

12.1.5 ML Fitting of Marginal Logistic Models: Other Methods

Alternative approaches have been proposed for ML fitting of marginal models. Lang and Agresti (1994) showed how to simultaneously fit a marginal model and an unsaturated loglinear model for π . For example, with binary matched pairs we can specify marginal logistic models for Y_1 in terms of x and for Y_2 in terms of x , and simultaneously model the log odds ratio between Y_1 and Y_2 in terms of x . The complete model can be specified as a special case of (12.2) and fitted using the constraint approach with Lagrange multipliers just described. In standard cases, the marginal and joint model parameters are orthogonal. If the marginal model holds, the ML estimator of the marginal model parameters is consistent even if the model for the joint distribution is incorrect.

Fitzmaurice and Laird (1993) proposed a related ML approach. A one-to-one correspondence holds between π and parameters of the saturated loglinear model. They used a further one-to-one correspondence between the main effect and the higher-order parameters of that loglinear model with the marginal probabilities and those same higher-order loglinear parameters. Models were then specified separately for the marginal probabilities and the higher-order (conditional) loglinear parameters. The likelihood is then maximized in terms of the two sets of model parameters. Again, the two sets of parameters are orthogonal, so the ML estimator of marginal model parameters is consistent when the marginal model holds. This *mixed parameter* approach is also available in specialized software (Kastner et al. 1997).

Yet another ML approach uses a one-to-one correspondence between π and parameters that describe the marginal distributions, the bivariate distributions, the trivariate distributions, and so on (e.g., Glonek and McCullagh 1995, Molenberghs and Lesaffre 1994, 1999). Multivariate logistic models then apply to the component distributions, although for simplicity some higher-order effects may be assumed to vanish. Glonek (1996) proposed a hybrid of this and the Fitzmaurice and Laird (1993) approach.

Finally, pseudo-likelihood approaches replace the likelihood function by a much simpler one, such as with the *composite likelihood* approach that uses contributions to the likelihood function for all pairs of observations. For an overview, see Varin et al. (2011).

12.2 MARGINAL MODELING: GENERALIZED ESTIMATING EQUATIONS (GEEs) APPROACH

For ML fitting of marginal models, at each combination of predictor values we assume a multinomial distribution over I^T cell probabilities for the T observations on an I -category response. As T increases, the number of multinomial probabilities increases dramatically. Currently, the ML fitting approaches described in the previous section are not practical when T is large or there are many predictors, especially when at least one is continuous.

An alternative to ML fitting uses a multivariate generalization of quasi-likelihood (Section 4.7). Recall that the (univariate) quasi-likelihood method, rather than assuming a particular distribution for Y , specifies only the first two moments; it links the mean to a linear predictor and also specifies how the variance depends on the mean. The estimates are solutions of estimating equations that are likelihood equations under the further assumption of a distribution in the exponential family with that mean and variance.

12.2.1 Generalized Estimating Equations Methodology: Basic Ideas

As in the univariate case, the quasi-likelihood method specifies a model for $\mu_t = E(Y_t)$ and specifies a variance function $v(\mu)$ that describes how $\text{var}(Y_t)$ depends on μ_t . Now, though, that model applies to the marginal distribution for each Y_t . The method also requires a working guess for the correlation structure among $\{Y_t\}$. The estimates are solutions of equations called *generalized estimating equations*. The method is often referred to as the GEE method. Liang and Zeger (1986) proposed it for marginal modeling with GLMs. We outline concepts here and give more technical details in Section 12.3.

In the GEE approach, we specify a variance function and a pairwise “working correlation” pattern for (Y_1, Y_2, \dots, Y_T) , but we do not need to assume a particular multivariate distribution. A popular working correlation structure is the *exchangeable* one that treats $\text{corr}(Y_t, Y_s)$ as identical for all s and t . For repeated measurement over time, also popular is the *autoregressive* structure, $\text{corr}(Y_t, Y_s) = \alpha^{|t-s|}$, which treats observations farther apart in time as more weakly correlated. More generally, an *unstructured* working correlation permits a separate correlation for each pair. In the other direction, a simple *independence* structure treats $\{Y_t\}$ as pairwise independent. However, the correlation is not the ideal parameter for describing association with categorical variables. Section 12.3.5 presents an adaptation of the GEE method based on choosing a working structure for the odds ratios to determine the relevant covariance matrix used in the generalized estimating equations.

The choice for the working correlation determines the GEE estimates of model parameters β describing effects of explanatory variables on $E(Y_t)$ and their model-based standard errors. For example, under the independence structure, the estimates are identical to the ML estimates obtained by treating all observations within and between clusters as independent. Usually, little a priori information is available about the correlation structure, and it is regarded as a nuisance. The GEE estimates of model parameters are valid, however, even if we misspecify the covariance structure. Specifically, suppose that the model is correct in the sense that the chosen link function and linear predictor truly describe how $E(Y_t)$ depend on the explanatory variables, $t = 1, \dots, T$. Then the GEE model parameter estimators are consistent (i.e., the estimators converge in probability to the true parameters). In practice, a chosen model is never exactly correct. This consistency result is useful, however, for suggesting that the correlation structure need not adversely affect this aspect of the estimates, for whatever model we use.

Although the model parameter estimates are usually fine whatever working correlation assumption we choose, their model-based standard errors are not. More appropriate standard errors result from an adjustment the GEE method can make using the empirical dependence the data exhibit. The standard errors based on the working correlation assumption are updated using the empirical dependence to yield more appropriate (*robust*) standard errors. So, even if we select a seemingly inappropriate working correlation structure such as independence, the empirical standard errors produced by the GEE method reflect the sample dependence.

Choosing a working correlation structure that well approximates the true correlations can pay benefits regarding efficiency of estimation of β . It may seem safest to use the unstructured correlation structure. When T is large, however, this approach can suffer some loss of efficiency because of the large number of correlation parameters that need to be estimated. Liang and Zeger (1986) noted that estimators based on independence working correlation can have surprisingly good efficiency when the actual correlation is weak to moderate. However, Fitzmaurice (1995) showed that efficiency can suffer for estimating the effect of an explanatory variable that varies within each cluster, especially when correlations between responses are moderately strong. To check the sensitivity to the selection or working correlation structure, we can compare results for different choices. In our experience, when the correlations are modest, all working correlation structures yield similar GEE estimates and

standard errors, as the empirical dependence has a large impact on adjusting the model-based standard errors. If they differed substantially, a more careful study of the correlation structure would be necessary. Unless we expect dramatic differences among the correlations, we recommend the exchangeable working correlation structure. This recognizes the dependence at the cost of only one extra parameter.

The GEE approach is appealing for categorical data because of its computational simplicity compared with ML. Advantages include not requiring specification of a joint distribution for (Y_1, Y_2, \dots, Y_T) , and the consistency of estimation even with misspecified correlation structure. However, it has limitations. Since the GEE approach does not completely specify the joint distribution, it does not have a likelihood function. Likelihood-based methods are not available for testing fit, comparing models, and conducting inference about parameters, as explained in Section 12.3.3. In fact, some statisticians (e.g., Lindsey and Lambert 1999) are critical of the GEE approach because of the lack of likelihood or possible conflicts between the nature of subject-specific effects and marginal effects. Others do not find this problematic, as they regard GEE as an estimation method rather than a model.

12.2.2 Example: Longitudinal Mental Depression Revisited

For [Table 12.1](#), comparing two treatments for mental depression, in Section 12.1.1 we used ML to fit a logistic model

$$\text{logit}[P(Y_t = 1)] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4(d \times t).$$

with drug \times time interaction. The GEE analysis provides similar results, regardless of the choice of working correlation structure. With the exchangeable structure, the estimated common correlation between pairs of the three responses is -0.003 . The successive observations apparently have pairwise appearance like independent observations. This is unusual for repeated measurement data. For this reason, similar results occur from fitting the model by treating the three observations for a subject as if they came from three separate subjects, that is, assuming $3 \times 340 = 1020$ independent observations rather than three correlated observations for each of 340 subjects.

[Table 12.5](#) shows results using the exchangeable working correlation structure. The empirical standard errors incorporate the sample dependence to adjust the exchangeable model-based standard errors. Here, there is not much change, since estimated correlations in the unstructured case do not vary much (0.07 for times 1 and 2, -0.03 for times 1 and 3, and -0.06 for times 2 and 3).

Table 12.5 Output from Using GEE to Fit Logistic Model to [Table 12.1](#), Using Exchangeable Working Correlation Structure

Analysis Of GEE Parameter Estimates					
Empirical Standard Error Estimates			Model-Based Standard Error Estimates		
Parameter	Estimate	Std Error	Parameter	Estimate	Std Error
Intercept	-0.0281	0.1742	Intercept	-0.0281	0.1638
severity	-1.3139	0.1460	severity	-1.3139	0.1459
drug	-0.0593	0.2286	drug	-0.0593	0.2222
time	0.4825	0.1199	time	0.4825	0.1150
drug*time	1.0172	0.1877	drug*time	1.0172	0.1891

The GEE estimated slope for the time effect (on the logit scale) for the standard drug is 0.4825, with empirical $SE = 0.1199$. For the new drug the slope increases by 1.0172, with empirical $SE = 0.1877$, thus giving strong evidence of a faster rate of improvement.

12.2.3 Example: Multinomial GEE Approach for Insomnia Trial

Liang and Zeger (1986) originally specified the GEE methodology for modeling univariate marginal distributions, such as the binomial and Poisson. It extends to marginal modeling of multinomial responses. Lipsitz et al. (1994) outlined a GEE approach, illustrating with cumulative logit and cumulative probit models. With this approach, for each pair of outcome categories we select working correlations for the pairs of repeated observations. Each multinomial response at a fixed observation uses the $(I - 1) \times (I - 1)$ multinomial covariance matrix.

We illustrate for the insomnia data of [Table 12.3](#), with Y_t = time to fall asleep with treatment x at occasion t . In Section 12.1.3 we fitted the marginal cumulative logit model [\(12.1\)](#), which is

$$\text{logit}[P(Y_t \leq j)] = \alpha_j + \beta_1 t + \beta_2 x + \beta_3(t \times x),$$

using ML. [Table 12.6](#) shows results using the GEE approach with independence working correlation structure (the only option with the SAS software employed). The GEE estimates are similar to the ML estimates from Section 12.1.3, and the same as the ML estimates we'd obtain by naively treating the pairs of responses as independent, that is, treating the data as 478 independent observations rather than as matched-pairs observations for 239 subjects. There is a positive association between the responses at the two times, for a given treatment, and the standard errors for the occasion effect for placebo (1.038) and the treatment-by-occasion interaction (0.708) are overestimated by treating the observations as independent. (Recall from Section 11.1.4 that positive dependence results in improved precision for estimating within-subject effects.) For the indicator coding used, the treatment effect (0.0336) refers to the initial time only; thus, it is based on two independent samples, and the empirical adjustment has essentially no effect.

Table 12.6 GEE Results for Marginal Model for Insomnia Data

Analysis Of GEE Parameter Estimates					
Empirical Standard Error Estimates			Model-Based Standard Error Estimates		
Parameter	Estimate	Std Error	Parameter	Estimate	Std Error
Intercept1	-2.2671	0.2188	Intercept1	-2.2671	0.2027
Intercept2	-0.9515	0.1809	Intercept2	-0.9515	0.1785
Intercept3	0.3517	0.1784	Intercept3	0.3517	0.1727
occasion	1.0381	0.1676	occasion	1.0381	0.2376
treat	0.0336	0.2384	treat	0.0336	0.2369
treat*occasion	0.7078	0.2435	treat*occasion	0.7078	0.3342

The substantive conclusions are the same as with ML fitting. Again, considerable evidence exists that the distribution of time to fall asleep decreased more over time for the treatment group than for the placebo group.

12.3 QUASI-LIKELIHOOD AND ITS GEE MULTIVARIATE EXTENSION: DETAILS

A GLM assumes a certain distribution for the response variable. Sometimes it is unclear how to select it. However, often there is a plausible relationship between the mean and variance, such as $v(\mu_i) = \phi\mu_i$ for count data. Then, an alternative to ML estimation is quasi-likelihood estimation (Section 4.7). We next present some details about this method and its GEE extension for marginal modeling of multivariate responses.

12.3.1 The Univariate Quasi-likelihood Method

We begin with models for a univariate response. For subject i , $i = 1, \dots, n$, let y_i be the outcome on Y with $\mu_i = E(Y_i)$ and variance function $v(\mu_i)$, and let x_{ij} be the value of explanatory variable j . For link function g , the linear predictor is $\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij} = \mathbf{x}_i^T \boldsymbol{\beta}$. The quasi-likelihood (QL) parameter estimates $\hat{\boldsymbol{\beta}}$ are the solutions of quasi-score equations

$$(12.3) \quad \mathbf{u}(\boldsymbol{\beta}) = \sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)^T v(\mu_i)^{-1} (y_i - \mu_i) = \mathbf{0},$$

where $\mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$. These *estimating equations* are the same as the likelihood equations (4.25) for GLMs when we substitute

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}.$$

They are not likelihood equations, however, without the extra assumption that $\{y_i\}$ has distribution in the natural exponential family. Under that assumption, $v(\mu_i)$ characterizes the distribution within the natural exponential family (Jørgensen 1987). Another motivation for equations (12.3) is that with $v(\mu_i)$ replaced by known variance v_i , they result from the weighted least-squares problem of minimizing $\sum_i [(y_i - \mu_i)^2 / v_i]$.

The likelihood equations (4.25) for a GLM depend only on the mean and variance of $\{y_i\}$ and the link function g , which determines $\partial \mu_i / \partial \eta_i$. Thus, Wedderburn (1974) suggested using them as estimating equations for *any* link and variance function, even if they do not correspond to a particular member of the natural exponential family.

12.3.2 Properties of Quasi-likelihood Estimators

In the quasi-likelihood (QL) method, the *quasi-score function* $u_j(\beta)$ in (12.3) is called an *unbiased estimating function*; this term refers to any function $h(y; \beta)$ of y and β such that $E[h(Y; \beta)] = 0$ for all β . The equations (12.3) that determine $\hat{\beta}$ are called *estimating equations*.

The quasi-likelihood method treats the quasi-score function as the derivative of a function called the *quasi-log likelihood*. This function may not be a proper log-likelihood function. Nonetheless, McCullagh (1983) showed that QL estimators have properties similar to those of ML estimators: Under correct specification of the mean and the variance function, they are asymptotically efficient among estimators that are locally linear in $\{y_i\}$. This result generalizes the Gauss–Markov theorem, although in an asymptotic rather than exact manner. The QL estimators $\hat{\beta}$ are asymptotically normal with model-based covariance matrix approximated by

$$(12.4) \quad V = \left[\sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial \beta} \right)^T [v(\mu_i)]^{-1} \left(\frac{\partial \mu_i}{\partial \beta} \right) \right]^{-1}.$$

This is equivalent to the formula for the large-sample covariance matrix of the ML estimator in a GLM [which is estimated by (4.31)].

A key result is that the QL estimator $\hat{\beta}$ is consistent for β (i.e., $\hat{\beta} \xrightarrow{P} \beta$) even if the variance function is misspecified, as long as the specification is correct for the link function and linear predictor. That is, assuming that the model form $g(\mu_i) = \sum_j \beta_j x_{ij}$ is correct, the consistency of $\hat{\beta}$ holds even if the true variance function is not $v(\mu_i)$. We now give a heuristic explanation for this.

When truly $\mu_i = g^{-1}(\sum_j \beta_j x_{ij})$, then from (12.3), $E[u_j(\beta)] = 0$ for all j . From (12.3), $u(\beta)/n$ is a vector of sample means. By a law of large numbers, it converges in probability to its expected value of $\mathbf{0}$. But solution $\hat{\beta}$ of the quasi-score equations is the value of β for which the sample mean is exactly equal to $\mathbf{0}$. Since $\hat{\beta}$ is a continuous function of these sample means, it converges to β . The consistency also follows from general results for unbiased estimating functions (see Note 12.4).

12.3.3 Sandwich Covariance Adjustment for Variance Misspecification

If we assume that $\text{var}(Y_i) = v(\mu_i)$ but the true $\text{var}(Y_i) \neq v(\mu_i)$, then the actual asymptotic covariance matrix of the QL estimator $\hat{\beta}$ is not V as given in (12.4). Instead, it is (Diggle et al. 2002, White 1982)

$$(12.5) \quad V \left[\sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial \beta} \right)^T [v(\mu_i)]^{-1} \text{var}(Y_i) [v(\mu_i)]^{-1} \left(\frac{\partial \mu_i}{\partial \beta} \right) \right] V.$$

Even though the variances are scalar, we express the matrices in this form to motivate the GEE multivariate extension discussed below. Matrix (12.5) simplifies to V if $\text{var}(Y_i) = v(\mu_i)$. In practice, the true variance function is unknown. A consistent estimator of (12.5) is a sample analog, replacing μ_i by $\hat{\mu}_i$ and $\text{var}(Y_i)$ by $(Y_i - \hat{\mu}_i)^2$ (Liang and Zeger 1986). The estimated covariance matrix is valid regardless of whether the variance specification $v(\mu_i)$ is correct. This estimated covariance is called a *sandwich estimator*, because the empirical evidence is sandwiched between the model-based covariance matrices.

To illustrate, for a sample of counts $\{y_i\}$, consider the common mean model, $\mu_i = \beta$, $i = 1, \dots, n$. Suppose we assume that $v(\mu_i) = \mu_i$, as in a Poisson model, but actually $\text{var}(Y_i) = \mu_i^2$. Since $\partial \mu_i / \partial \beta = 1$, from (12.3),

$$u(\beta) = \sum_i \left(\frac{\partial \mu_i}{\partial \beta} \right) v(\mu_i)^{-1} (y_i - \mu_i) = \sum_i \frac{(y_i - \mu_i)}{\mu_i} = \sum_i \frac{(y_i - \beta)}{\beta}.$$

Setting this equal to 0 and solving, $\hat{\beta} = (\sum_i y_i)/n = \bar{y}$. So, the model-based variance (12.4) simplifies to

$$V = \left[\sum_i \left(\frac{\partial \mu_i}{\partial \beta} \right) [v(\mu_i)]^{-1} \left(\frac{\partial \mu_i}{\partial \beta} \right) \right]^{-1} = \left[\sum_i \mu_i^{-1} \right]^{-1} = \frac{\beta}{n}.$$

The actual asymptotic variance (12.5) of $\hat{\beta} = \bar{y}$ that takes into account the variance misspecification is

$$V \left[\sum_i \left(\frac{\partial \mu_i}{\partial \beta} \right) [v(\mu_i)]^{-1} \text{var}(Y_i) [v(\mu_i)]^{-1} \left(\frac{\partial \mu_i}{\partial \beta} \right) \right] V = \frac{\beta}{n} \left[\sum_i \mu_i^{-1} \mu_i^2 \mu_i^{-1} \right] \frac{\beta}{n} = \frac{\beta^2}{n}.$$

In practice, we replace the true variance μ_i^2 in this expression by $(y_i - \bar{y})^2$, so the last expression simplifies (using $\mu_i = \beta$) to $\sum_i (y_i - \bar{y})^2/n^2$, a quite sensible estimate of the variance of a sample mean. Using this sandwich estimator instead of $V = \hat{\beta}/n = \bar{y}/n$ protects against an incorrect choice of variance function.

The purpose of the sandwich estimator is to use the data's empirical evidence about variation to adjust the standard errors in case the true variance differs substantially from the working guess. Inference uses the asymptotic normality of the estimators together with the sandwich-estimated covariance matrix. A significance test of $H_0: \beta_j = 0$ using test statistic $z = \hat{\beta}_j/SE$ (or its square) and a 95% confidence interval $\hat{\beta}_j \pm 1.96(SE)$ provide Wald-type inference.

In summary, even with incorrect specification of the variance function, we can still consistently estimate β and estimate the asymptotic covariance of $\hat{\beta}$ by estimating the sandwich adjustment (12.5). However, some efficiency loss occurs when the variance chosen, $v(\mu_i)$, is wildly inaccurate. Also, the number of clusters n may need to be large for the sample version of (12.5) to work well; otherwise, the empirically based standard errors tend to underestimate the true ones (e.g., Firth 1993b). As estimators, those standard errors can also show more variability than parametric estimators (Kauermann and Carroll 2001). Boos (1992) proposed analogs of score tests that solve the GEE under the restriction that the null holds and which incorporate empirical variance estimates, illustrating with tests for trend and lack of fit in binary regression. See also Lefkopoulou et al. (1989) and Rotnitzky and Jewell (1990). Finally, in practice we must recognize that just as the variance function chosen only approximates the true one, so is the specification for the mean only approximate.

12.3.4 GEE Multivariate Methodology: Technical Details

Now we consider the generalized estimating equations (GEE) multivariate version of QL. For cluster i , let $\mathbf{y}_i = (y_{i1}, \dots, y_{iT_i})^T$ and $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{iT_i})^T$, where $\mu_{it} = E(Y_{it})$. The number T_i of responses may vary by cluster. Let \mathbf{x}_{it} denote a $p \times 1$ vector of explanatory variable values for y_{it} . The notation allows also values of the explanatory variables to vary for the observations in a cluster. The linear predictor of the model is $\eta_{it} = g(\mu_{it}) = \mathbf{x}_{it}^T \boldsymbol{\beta}$ for link function g . The model refers to the marginal distribution at each t rather than the joint distribution. Let \mathbf{X}_i be the $T_i \times p$ matrix of explanatory variable values for cluster i , for which row t is \mathbf{x}_{it}^T .

We assume that y_{it} has probability mass function of form

$$f(y_{it}; \theta_{it}, \phi) = \exp\{[y_{it}\theta_{it} - b(\theta_{it})]/\phi + c(y_{it}, \phi)\}.$$

When ϕ is known, this is the natural exponential family with natural parameter θ_{it} . From Section 4.4.2,

$$\mu_{it} = E(Y_{it}) = b'(\theta_{it}), \quad v(\mu_{it}) = \text{var}(Y_{it}) = b''(\theta_{it})\phi.$$

The GEE method also assumes a working correlation matrix $\mathbf{R}(\boldsymbol{\alpha})$ for \mathbf{Y}_i , depending on parameters $\boldsymbol{\alpha}$. The exchangeable working correlation has $\text{corr}(Y_{it}, Y_{is}) = \alpha$ for each pair in \mathbf{Y}_i . Let $\mathbf{b}_i(\boldsymbol{\theta}) = (b(\theta_{i1}), \dots, b(\theta_{iT_i}))$, and let \mathbf{B}_i denote a diagonal matrix with main-diagonal elements $\mathbf{b}_i''(\boldsymbol{\theta})$. Then the working covariance matrix for \mathbf{Y}_i is

$$(12.6) \quad \mathbf{V}_i = \mathbf{B}_i^{1/2} \mathbf{R}(\boldsymbol{\alpha}) \mathbf{B}_i^{1/2} \phi.$$

If \mathbf{R} is the true correlation matrix for \mathbf{Y}_i , then $\mathbf{V}_i = \text{cov}(\mathbf{Y}_i)$.

Let Δ_i be the diagonal matrix with elements $\partial\theta_{it}/\partial\eta_{it}$ on the main diagonal for $t = 1, \dots, T_i$. (For the canonical link, this is the identity matrix.) Let $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta} = \mathbf{B}_i \Delta_i \mathbf{X}_i$ be a $T_i \times p$ matrix with typical element expressing $\partial\mu_{it}/\partial\beta_j$ in the form $(\partial\mu_{it}/\partial\theta_{it})(\partial\theta_{it}/\partial\eta_{it})(\partial\eta_{it}/\partial\beta_j)$. From (12.3), for univariate GLMs the quasi-likelihood estimating equations have the form

$$\sum_{i=1}^n (\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta})^T v(\mu_i)^{-1} [\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] = \mathbf{0},$$

where $\mu_i = \mu_i(\boldsymbol{\beta}) = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$. The analog of this in the multivariate case is the set of *generalized estimating equations*

$$(12.7) \quad \sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} [\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] = \mathbf{0}.$$

The GEE estimator $\hat{\boldsymbol{\beta}}$ is the solution of these equations.

The approach that sets $\mathbf{R}(\boldsymbol{\alpha}) = \mathbf{I}$ treats clustered responses as independent. In that case, (12.6) simplifies to $\mathbf{V}_i = \mathbf{B}_i \phi$, and the generalized estimating equations simplify to

$$\begin{aligned} \sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} [\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] &= \sum_{i=1}^n \mathbf{X}_i^T \Delta_i \mathbf{B}_i \mathbf{V}_i^{-1} [\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] \\ &= (1/\phi) \sum_{i=1}^n \mathbf{X}_i^T \Delta_i [\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] = \mathbf{0}. \end{aligned}$$

The solution $\hat{\boldsymbol{\beta}}$ is then the same as the ordinary ML estimator for a GLM with the chosen link function and variance function, treating $(y_{i1}, \dots, y_{iT_i})$ as independent observations.

Normally, we select a working correlation matrix permitting dependence, such as the exchangeable structure. Liang and Zeger (1986) suggested computing the GEE estimates by iterating between a modified Fisher scoring algorithm for solving the generalized estimating equations for $\boldsymbol{\beta}$ (given current estimates of $\boldsymbol{\alpha}$ and ϕ) and using residuals for moment estimation of $\boldsymbol{\alpha}$ and ϕ (based on the current estimates of $\boldsymbol{\beta}$). They suggested estimates of $\mathbf{R}(\boldsymbol{\alpha})$ for a variety of correlation structures. Alternative algorithms simultaneously solve estimating equations for $\boldsymbol{\beta}$ and for association

parameters. See Liang et al. (1992) and Note 12.5.

Liang and Zeger (1986) showed asymptotic normality and consistency as the number of clusters n increases. Under certain regularity conditions including appropriate consistency for estimates of α and φ ,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(\mathbf{0}, V_G).$$

Here, generalizing (12.5), $V_G = \lim_{n \rightarrow \infty} V_{G,n}$ with

$$V_{G,n} = n \left[\sum_{i=1}^n D_i^T V_i^{-1} D_i \right]^{-1} \left[\sum_{i=1}^n D_i^T V_i^{-1} \text{cov}(Y_i) V_i^{-1} D_i \right] \left[\sum_{i=1}^n D_i^T V_i^{-1} D_i \right]^{-1}.$$

The estimated sandwich covariance matrix $(1/n)V_{G,n}$ of $\hat{\beta}$ replaces β with $\hat{\beta}$, φ with $\hat{\varphi}$, α with $\hat{\alpha}$, and $\text{cov}(Y_i)$ by $[\mathbf{y}_i - \mu_i(\hat{\beta})][\mathbf{y}_i - \mu_i(\hat{\beta})]^T$.

When the working correlation structure is the true one and $\text{cov}(Y_i) = V_i$, the asymptotic covariance matrix $(1/n)V_{G,n}$ simplifies to the model-based covariance matrix, $(\sum_i D_i^T V_i^{-1} D_i)^{-1}$. This is the relevant covariance matrix if we put complete faith in our choice of the correlation structure.

12.3.5 Working Associations Characterized by Odds Ratios

With binary data, the correlation is not the best way to characterize within-cluster association. The marginal probabilities constrain the possible correlation values, since the range of possible values for $E(Y_{it}Y_{is}) = P(Y_{it} = 1, Y_{is} = 1)$ depends on $P(Y_{it} = 1)$ and $P(Y_{is} = 1)$. In particular, it is not possible to have a strong correlation when the marginal probabilities vary greatly. Also, the correlations typically depend on values of explanatory variables. Although a structure such as exchangeable correlations may be plausible for underlying continuous latent variables, it is usually not so for observed binary response variables. In some cases, working correlation matrices may not have valid joint distributions with that structure, and there can be a breakdown in the asymptotic properties of the correlation and model parameter estimators; see Chaganty and Joe (2004, 2006), Crowder (1995), and Touloumis(2011).

An alternative approach uses the odds ratio to characterize pairwise associations. For instance, we can model the log odds ratios for pairs of observations in a cluster as exchangeable, and use the odds ratios together with the marginal probabilities to specify working correlation matrices. This approach also has the advantages that the association parameters are distinct from the means and that the working correlations can depend on values of explanatory variables. See Fitzmaurice et al. (1993) and Lipsitz et al. (1991).

Carey et al. (1993) suggested an iterative *alternating logistic regressions* algorithm. It alternates between a GEE step for the regression parameters in the model for the mean and a step for an association model for the log odds ratio. This is especially useful when the structure of the association is itself a major focus rather than a nuisance, as the method also provides standard errors for the estimates for the association model.

We illustrate for the GEE analysis of the depression study of [Table 12.1](#). Using an exchangeable odds ratio for pairs of times, we obtain a common log odds ratio estimate of -0.007 with $SE = 0.162$. The parameter estimates and standard errors are the same to three decimal places as in the correlation-based analysis of Section 12.2.2.

12.3.6 GEE Approach: Multinomial Responses

Lipsitz et al. (1994) developed the GEE approach for marginal modeling with a multinomial response. Let $y_{it}(j) = 1$ if observation t in cluster i has outcome j ($j = 1, \dots, I - 1$). Let \mathbf{y}_i be the $T_i(I - 1)$ binary indicators for cluster i . Then, the chosen $[T_i(I - 1)] \times [T_i(I - 1)]$ working covariance matrix \mathbf{V}_i for \mathbf{y}_i specifies a pattern for $\text{corr}(Y_{it}(j), Y_{is}(k))$ for each pair of outcome categories (j, k) and each pair (t, s) . The $(I - 1) \times (I - 1)$ block of \mathbf{V}_{it} for $(y_{it}(1), \dots, y_{it}(I - 1))$ is a multinomial covariance matrix with $v_{it}(j) = P(Y_{it}(j) = 1)[1 - P(Y_{it}(j) = 1)]$ on the main diagonal and $-P(Y_{it}(j) = 1)P(Y_{it}(k) = 1)$ off it. The remaining elements of \mathbf{V}_i contain elements $\text{cov}(Y_{it}(j), Y_{is}(k))$. For instance, one possibility is the exchangeable structure, $\text{con}(Y_{it}(j), Y_{is}(k)) = \rho_{jk}$ for all t and s . An alternative approach specifies working associations based on multinomial odds ratios such as the set of local odds ratios (Touloumis 2011).

The generalized estimating equations for β again have the form

$$\mathbf{u}(\beta) = \sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0},$$

where $\boldsymbol{\mu}_i$ is the vector of probabilities associated with \mathbf{y}_i , and $\mathbf{D}_i^T = \partial \boldsymbol{\mu}_i^T / \partial \beta$. Lipsitz et al. (1994) suggested a Fisher scoring algorithm for solving these equations and a method of moments update for estimating $\{\rho_{jk}\}$ at each step of the iteration. An empirically adjusted sandwich covariance matrix of $\hat{\beta}$ is again

$$\left[\sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i \right]^{-1} \left[\sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} \text{cov}(\mathbf{Y}_i) \mathbf{V}_i^{-1} \mathbf{D}_i \right] \left[\sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i \right]^{-1}.$$

This is estimated by substituting $\hat{\boldsymbol{\mu}}_i$ from the model fit and replacing $\text{cov}(\mathbf{Y}_i)$ by the empirical covariance matrix of \mathbf{y}_i .

12.3.7 Dealing with Missing Data

Unfortunately, studies with repeated measurements often have cases for which at least one response in a cluster is missing. In a longitudinal study, for instance, some subjects may drop out before its conclusion. This often results in a *monotone missingness* pattern, in which if y_t was observed then so necessarily was y_{t-1} , and if y_t was missing then so necessarily was y_{t+1} . When data are missing, biased estimates can result from either analyzing the observed data as if no data are missing or from analyzing the data file that deletes entire clusters that have at least some missingness.

We partition the data \mathbf{Y} into those that are observed, $\mathbf{Y}^{(o)}$, and those that are missing, $\mathbf{Y}^{(m)}$. Let \mathbf{M} denote a vector of missing data indicators that equal 1 when an observation is missing and 0 otherwise. Little and Rubin (2002) called the data *missing completely at random* (MCAR) if \mathbf{M} is statistically independent of \mathbf{Y} ; that is, the probability that an observation is missing is independent of that observation's value and the values of other variables in the data set. In this case, $\mathbf{Y}^{(m)}$ behaves like a random sample from \mathbf{Y} . Less restrictively, they called the data *missing at random* (MAR) if the distribution of $(\mathbf{M}|\mathbf{Y})$ equals that of $(\mathbf{M}|\mathbf{Y}^{(o)})$; that is, what caused the data to be missing does not depend on the data itself. For example, in a longitudinal study, if whether someone drops out of the study depends on values observed prior to the drop-out but not the later unobserved values, the data are MAR.

Consider a study such as in Section 12.1.1, modeling depression as a function of treatment, severity, and time. If the probability that the depression assessment is missing is the same for all subjects regardless of treatment, severity, and time, then the data are MCAR. If the probability the depression assessment is missing varies according to time but does not vary according to the depression assessment of subjects at the same treatment, severity, and time, then the data are not MCAR but are MAR. They are not MAR if those with a missing depression assessment tend to have worse depression than those not missing the assessment, controlling for the other variables.

In practice, we cannot test whether MCAR or MAR is satisfied, because we do not know the values of the missing data. However, certain evidence can show that they are not satisfied. For example, suppose the subjects classified as severe in their depression symptoms tended to be much more likely to have missing depression observations than those classified as mild. Then, the missing data do not seem to be MCAR, because the missing observations do not resemble a random sample of all the observations.

When either MCAR or MAR is plausible, with a likelihood-based analysis it is not necessary to model the missingness mechanism. An analysis using only $\mathbf{Y}^{(o)}$ is not systematically biased. For ML fitting, we treat the missing data as random variables to be integrated out of the likelihood function using the EM algorithm (Section 13.6.3). Exercises 12.12 and 12.29 illustrate how this can be done simply for a contingency table having monotone missingness. An alternative to ML uses *multiple imputation*. This is a Monte Carlo method in which the missing values are replaced several times by simulated versions from their conditional distribution, given the observed data (see Rubin 1996). Each simulated complete data set is analyzed with standard methods, and the results are then combined. The resulting estimates have standard errors based on the within-imputation and between-imputation variances, thus incorporating the missing-data uncertainty. This method is most naturally applied in a Bayesian context that also treats parameters as random.

With the GEE method, different clusters can have different numbers of observations. The data input file has a separate line for each observation, and for longitudinal studies, computations use those times for which a subject has an observation. However, bias can arise in GEE estimates when data are missing unless the data are MCAR. The missingness can then be ignored and an analysis using the observed data only is valid. In the MAR case, it is valid when estimating equations can be weighted by response probabilities (Robins et al. 1995). Otherwise, however, with GEE and other non-likelihood-based methods, the missingness process cannot be ignored even in the MAR case. Kenward et al. (1994) illustrated the potential breakdown in GEE estimates when the data are not MCAR.

Often, missingness is not MCAR or MAR but rather is *informative* and cannot be ignored. For instance, in a longitudinal study measuring pain, perhaps a subject dropped out when the pain got above some threshold. Then, more complex analyses are needed that model the missingness as well as the complete data. That is, methods require a joint distribution for \mathbf{Y} and \mathbf{M} (Little 2005). Let $f(\cdot)$ denote a generic probability mass function, which also depends on explanatory variables \mathbf{x} and parameters. *Selection models* factor the joint distribution of \mathbf{Y} and \mathbf{M} as

$$f(\mathbf{y}, \mathbf{M}; \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\psi}) = f(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta})f(\mathbf{M}|\mathbf{y}; \mathbf{x}, \boldsymbol{\psi}),$$

where $f(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta})$ is the model in the absence of missing values and $f(\mathbf{M}|\mathbf{y}; \mathbf{x}, \boldsymbol{\psi})$ is the model for the missing-data mechanism, such as a logistic model for \mathbf{M} that selects dropouts according to their history of previous responses. *Pattern mixture models* use the alternative factorization,

$$f(\mathbf{y}, \mathbf{M}; \mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) = f(\mathbf{y}|\mathbf{M}, \mathbf{x}, \boldsymbol{\phi})f(\mathbf{M}; \mathbf{x}, \boldsymbol{\theta}),$$

which conditions the distribution of \mathbf{Y} on the missing data pattern. The two specifications are equivalent when \mathbf{M} is independent of \mathbf{Y} , with $\boldsymbol{\beta} = \boldsymbol{\phi}$ and $\boldsymbol{\psi} = \boldsymbol{\theta}$. For discussion of advantages of each modeling approach and details on ways of modeling missingness, see Little (2005) and references in Note 12.6.

Analyses in the presence of much missingness should be made with caution. Typically, little is known about the missing data mechanism, and assumptions about it cannot be checked. Since inferences may not be robust, a sensitivity study is necessary to check how results depend on specification of that mechanism. In the absence of a model for the missingness, we should at least compare results of the analysis using all available cases for all clusters to the analysis using only clusters having no missing observations. If results differ substantially, conclusions should be very tentative until the reasons for missingness can be studied.

12.4 TRANSITIONAL MODELS: MARKOV CHAIN AND TIME SERIES MODELS

For responses Y_t , $t = 0, 1, 2, \dots$, the indexed family of random variables $(Y_0, Y_1, Y_2, Y_3, \dots)$ is called a *stochastic process*. Its *state space* is the set of possible values for Y_t . When the state space is categorical, $\{Y_t\}$ has *discrete state space*. Often, the main focus is on the dependence of Y_t on the other responses as well as any explanatory variables. Models of this type are called *transitional models*, as they describe the transition to the state for Y_t from the states at other observations. When t is a time index and the dependence of Y_t on other responses is modeled solely in terms of responses $\{y_0, y_1, \dots, y_{t-1}\}$ observed previously, the model is referred to as a *time series* model.

Let $f(y_0, \dots, y_T)$ denote the joint probability mass function of (Y_0, \dots, Y_T) (ignoring, for now, explanatory variables). Transitional models for time series data use the factorization

$$f(y_0, \dots, y_T) = f(y_0)f(y_1|y_0)f(y_2|y_0, y_1)\cdots f(y_T|y_0, y_1, \dots, y_{T-1}).$$

Unlike the marginal models in the other sections of this chapter, this modeling is conditional on previous responses.

12.4.1 Markov Chains

A *Markov chain* is a simple stochastic process having discrete state space for which, for all t , the conditional distribution of Y_t , given Y_0, \dots, Y_{t-1} , is identical to the conditional distribution of Y_t given Y_{t-1} alone. That is, given Y_{t-1} , Y_t is conditionally independent of Y_0, \dots, Y_{t-2} . Knowing the present state of a Markov chain, information about past states does not help us predict future states. For Markov chains,

$$(12.8) \quad f(y_0, \dots, y_T) = f(y_0)f(y_1|y_0)f(y_2|y_1)\cdots f(y_T|y_{T-1}).$$

Many transitional models have Markov chain structure.

For a Markov chain, denote the conditional probability $P(Y_t = j|Y_{t-1} = i)$ by $\pi_{j|i}(t)$. The $\{\pi_{j|i}(t)\}$, which satisfy $\sum_j \pi_{j|i}(t) = 1$, are called *transition probabilities*. The $I \times I$ matrix $\{\pi_{j|i}(t), i = 1, \dots, I, j = 1, \dots, I\}$ is a *transition probability matrix*. From (12.8), the joint distribution for a Markov chain depends only on one-step transition probabilities and the marginal distribution for the initial state Y_0 . It follows that the joint distribution satisfies loglinear model

$$(Y_0Y_1, Y_1Y_2, \dots, Y_{T-1}Y_T).$$

For a sample of realizations of a discrete-time Markov chain, a contingency table displays counts of the possible sequences. A test of fit of this loglinear model checks whether the process plausibly satisfies the Markov property.

Statistical inference for Markov chains uses standard methods of categorical data analysis. For example, consider ML estimation of transition probabilities. Let $n_{ij}(t)$ denote the number of transitions from state i at time $t - 1$ to state j at time t . For fixed t , $\{n_{ij}(t)\}$ form the two-way marginal table for dimensions $t - 1$ and t of an I^{T+1} contingency table. For the $n_{i+}(t)$ subjects in category i at time $t - 1$, suppose that $\{n_{ij}(t), j = 1, \dots, I\}$ have a multinomial distribution with parameters $\{\pi_{j|i}(t)\}$. Let $\{n_{i0}\}$ denote the initial counts. Suppose that they also have a multinomial distribution, with parameters $\{\pi_{i0}\}$. If subjects behave independently, from (12.8) the likelihood function is proportional to

$$(12.9) \quad \left(\prod_{i=1}^I \pi_{i0}^{n_{i0}} \right) \left\{ \prod_{t=1}^T \prod_{i=1}^I \left[\prod_{j=1}^I \pi_{j|i}(t)^{n_{ij}(t)} \right] \right\}.$$

The transition probabilities are parameters of IT independent multinomial distributions. From Anderson and Goodman (1957), the ML estimates are

$$\hat{\pi}_{j|i}(t) = n_{ij}(t)/n_{i+}(t).$$

Many models assume that the transition probabilities are *stationary*: For all i and j ,

$$\pi_{j|i}(1) = \pi_{j|i}(2) = \cdots = \pi_{j|i}(T) = \pi_{j|i}.$$

Let $n_{ij} = \sum_t n_{ij}(t)$. Under the assumption of stationary transition probabilities, the likelihood in (12.9) simplifies, and the ML estimators are

$$\hat{\pi}_{j|i} = n_{ij}/n_{i+}.$$

These results generalize to more complex dependences. For example, a stochastic process is a *kth-order Markov chain* if, for all t , the conditional distribution of Y_t , given Y_0, \dots, Y_{t-1} , is identical to the conditional distribution of Y_t , given $(Y_{t-1}, \dots, Y_{t-k})$. Given the states at the previous k times, the future behavior of the chain is independent of past behavior before those k times. The Markov chain as defined above is first-order.

12.4.2 Example: Changes in Evapotranspiration Rates

Fokianos and Kedem (2002, p. 39) analyzed a time series that consists of a series of 84 indicators about monthly changes in evapotranspiration (evaporation plus transpiration) rates, compared with a year earlier, at a location in southern Israel. At a particular time t , $y_t = 1$ if the change from a year ago is greater than the average, and $y_t = 0$ if it less than average. In order of the 84 months, they presented the data:

1 1 1 1 1 1 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 1 0 0 0 1 0 0 0 0 1 1 1 1 1

1 1 0 0 1 1 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 0

The logistic regression model with the autoregressive structure

$$\text{logit}[P(Y_t = 1|y_{t-1}, y_{t-2}, \dots, y_0)] = \alpha + \beta_1 y_{t-1}$$

is a first-order Markov chain with stationary transition probabilities. A second-order Markov chain with stationary transition probabilities has

$$\text{logit}[P(Y_t = 1|y_{t-1}, y_{t-2}, \dots, y_0)] = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2}.$$

If the second-order model holds, it is sufficient to fit the model to the data as summarized in [Table 12.7](#), treating the data as four independent binomial variates. Using standard logistic software, we get a deviance of 1.47 (df = 1), and estimates $(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2) = (-1.439, 3.970, -1.304)$ with SE values (0.427, 1.078, 1.078). This suggests that the simpler first-order model may be adequate. Fitting the model with constraint $\beta_2 = 0$ increases the deviance by 1.98, with $(\hat{\alpha}, \hat{\beta}_1) = (-1.609, 2.996)$ having SE values (0.414, 0.572). Here, $\hat{\beta}_1$ necessarily equals the log odds ratio for the collapsed 2×2 table relating y_t and Y_{t-1} . The interpretation is that the estimated odds that $y_t = 1$ is $\exp(2.996) = 20$ times as high when $Y_{t-1} = 1$ as it is when $Y_{t-1} = 0$. For the first-order model, we can obtain an extra observation by considering the pair at times 0 and 1; then, $\hat{\beta}_2$ changes to 3.027. Fitting the model permitting dependence higher than second-order provides little improvement in the deviance.

Table 12.7 Summary of Two-Step Transitions for Evapotranspiration Data

Y_{t-1}	Y_{t-2}	Y_t	
		1	0
1	1	26	7
	0	6	1
0	1	0	8
	0	7	27

12.4.3 Transitional Models with Explanatory Variables

Transitional models can also include explanatory variables \mathbf{x} . The joint mass function of T sequential responses is then

$$\begin{aligned} f(y_1, \dots, y_T; \mathbf{x}) \\ = f(y_1; \mathbf{x})f(y_2|y_1; \mathbf{x})f(y_3|y_1, y_2; \mathbf{x}) \cdots f(y_T|y_1, y_2, \dots, y_{T-1}; \mathbf{x}). \end{aligned}$$

More generally, \mathbf{x} may take a different value for each component, such as when covariates are time dependent.

With binary y , we can specify a logistic regression model for each term in this factorization. An example is the k th-order model

$$\begin{aligned} f(y_t|y_1, \dots, y_{t-1}; \mathbf{x}_t) \\ = \frac{\exp[y_t(\alpha + \beta_1 y_{t-1} + \dots + \beta_k y_{t-k} + \boldsymbol{\beta}^T \mathbf{x}_t)]}{1 + \exp(\alpha + \beta_1 y_{t-1} + \dots + \beta_k y_{t-k} + \boldsymbol{\beta}^T \mathbf{x}_t)}, \quad y_t = 0, 1, \end{aligned}$$

which also treats k previous responses as explanatory variables. It is called a *regressive logistic model* (Bonney 1987). In the special case of first-order Markov structure, the coefficients of $\{y_1, \dots, Y_{t-2}\}$ equal 0 in the model for y_t (Azzalini 1994, Bonney 1987). It may also help to allow interaction between \mathbf{x}_t and Y_{t-1} in their effects on y_t .

The interpretation and magnitude of $\boldsymbol{\beta}$ depends on how many previous observations are in the model. Normally we'd use the same number k of previous observations in each term. Within-cluster effects may diminish markedly by conditioning on previous responses. This is an important difference from marginal models, for which the interpretation does not depend on the specification of the dependence structure.

For a given subject, the product of the conditional mass functions over t determines that subject's contribution to the likelihood function. It is common to ignore the contribution of the marginal distribution for the first term. So, given the predictor, model-fitting treats repeated transitions by a subject as independent. Thus, we can fit the model with ordinary GLM software, treating each transition as a separate observation. Even when \mathbf{x} is also treated as random, Fokianos and Kedem (2002) formed a likelihood function using only this product of conditional mass functions for $\{y_t\}$. They termed it a *partial likelihood*, showed theoretical properties, and argued that relatively little information is lost using it.

More generally, Fahrmeir and Kaufmann (1987) and Fokianos and Kedem (2002, 2003) proposed a generalized linear model structure in which a link function applied to $E(Y_t)$ is modeled as a linear function of past observations and random time-dependent explanatory variables, without assuming Markov structure or stationarity. Moreover, Y_t can be multivariate, such as a set of $I - 1$ indicators for a I -category multinomial response. See Note 12.8 and Section 13.3.9 for other approaches for categorical time series data.

12.4.4 Example: Child's Respiratory Illness and Maternal Smoking

[Table 12.8](#) is from a Harvard study of air pollution and health. At ages 7 through 10, children were evaluated annually on the presence of respiratory illness. A predictor is maternal smoking at the start of the study, where $s = 1$ for smoking regularly and $s = 0$ otherwise. Let y_t denote the response at age t ($t = 7, 8, 9, 10$). We use the first-order regressive logistic model

Table 12.8 Child's Respiratory Illness by Age and Maternal Smoking

Child's Respiratory Illness			No Maternal Smoking		Maternal Smoking		
Age 7	Age 8	Age 9	Age 10		Age 10		
			No	Yes	No	Yes	
No	No	No	237	10	118	6	
		Yes	15	4	8	2	
	Yes	No	16	2	11	1	
		Yes	7	3	6	4	
	Yes	No	24	3	7	3	
		Yes	3	2	3	1	
			No	6	4	2	
			Yes	5	11	7	

Source: Data courtesy of James Ware.

$$\text{logit}[P(Y_t = 1)] = \alpha + \beta_1 y_{t-1} + \beta_2 s + \beta_3 t, \quad t = 8, 9, 10.$$

Each subject contributes three observations to the model fitting. The data set consists of 12 binomials, for the $2 \times 2 \times 3$ combinations of (Y_{t-1}, s, t) . For instance, for the combination $(0, 0, 8)$, $y_8 = 0$ for $237 + 10 + 15 + 4 = 266$ subjects and $y_8 = 1$ for $16 + 2 + 7 + 3 = 28$ subjects. The ML fit is

$$\text{logit}[\hat{P}(Y_t = 1)] = -0.293 + 2.211 y_{t-1} + 0.296 s - 0.243 t,$$

with effect SE values (0.158, 0.156, 0.095). Not surprisingly, the previous observation has a strong effect. Given that and the child's age, there is slight evidence of a positive effect of maternal smoking: The likelihood-ratio statistic for $H_0: \beta_2 = 0$ is 3.55 (df = 1, $P = 0.06$). The model itself does not show any evidence of lack of fit ($G^2 = 3.12$, df = 8).

12.4.5 Example: Initial Response in Matched Pair as a Covariate

Consider matched-pairs data in which the observations occur at different times. It can be more relevant to model the follow-up response using the initial response as a covariate, rather than treating the two variables symmetrically in a marginal model.

We illustrate with the insomnia study of [Table 12.3](#) from Section 12.1.3. Let Y_2 denote the follow-up ordinal response, for treatment x with initial observation y_1 . In the transitional model

$$(12.10) \text{ logit}[P(Y_2 \leq j)] = \alpha_j + \beta_1 x + \beta_2 y_1,$$

β_1 compares the follow-up distributions for the treatments, adjusting for the initial observation. This is an analog of an analysis-of-covariance model, with ordinal rather than continuous response. This cumulative logit model refers to a univariate response (Y_2) rather than marginal distributions of a multivariate response (Y_1, Y_2).

In this model, we use scores (10, 25, 45, 75) for the four categories of the initial time to fall asleep y_1 . Applying software for ordinary cumulative logit models to the univariate response Y_2 , the ML treatment effect estimate is $\hat{\beta}_1 = 0.885$ ($SE = 0.246$). This provides strong evidence that follow-up time to fall asleep is lower for the active drug group. For any given value for the initial response, the estimated odds of falling asleep by a particular time for the active treatment are $\exp(0.885) = 2.42$ times those for the placebo group.

For matched-pairs data, marginal models evaluate how the marginal distributions of Y_1 and Y_2 depend on explanatory variables. By contrast, a transitional model treats Y_2 as a univariate response, evaluating effects of explanatory variables while adjusting for the initial response y_1 . In some situations, whether an effect of a certain type exists may differ between these two types of model. For example, suppose the true marginal distributions for initial response are identical for the treatment groups, as we expect with random assignment of subjects to the groups. Suppose also that there is no treatment effect, in the sense that conditional on the initial response, the follow-up response distribution is identical for the treatment groups. Then, the follow-up marginal distributions are also identical. By contrast, suppose the initial marginal distributions are not identical, as might well happen with observational data for which randomization of subjects is not possible. Then, even when the conditional distributions for follow-up response are identical for the two treatment groups, the difference between follow-up and initial marginal distributions may differ between the treatment groups. In such cases, it is more informative to use the transitional model.

12.4.6 Transitional Models and Loglinear Conditional Models

In this chapter we have mainly focused on marginal models. Transitional models, by contrast, are *conditional*, with the effects on y_t being described conditionally on previously observed responses.

More generally, we could ignore the ordering on t and consider models in which y_t is modeled in terms of y_u with both $u < t$ and with $u > t$. This is essentially what we do with standard loglinear models, in which we model effects, associations, and interactions conditional on all the other response variables. Although such models are of use for describing joint distributions of several response variables, as we did in Chapters 9 and 10, they are usually of less relevance when we are analyzing effects of explanatory variables \mathbf{x} . Normally, in considering the effect of an explanatory variable on a response y_t , it is not relevant to describe that effect conditional on all other responses.

NOTES

Section 12.1: Marginal Modeling: Maximum Likelihood Approach

12.1 Marginal ML: For other work on ML fitting of marginal models, see Bergsma and Rudas (2002), Bergsma et al. (2009), Colombi (1998), Drton and Richardson (2008), Ekholm et al. (2000), Fitzmaurice and Laird (1993), Lang (2004, 2005), Lang et al. (1999), and Molenberghs and Verbeke (2005, Chap. 6, 7), with the last reference also modeling global odds ratios for ordinal responses. Ekholm et al. (2000) modeled association factors (Sec. 2.4.2, which they referred to as *dependence ratios*) and corresponding higher-order measures that they used together with the marginal probabilities to parameterize a multivariate binary response. Ashford and Sowden (1970), Lesaffre and Molenberghs (1991), and Ochi and Prentice (1984) presented multivariate probit models that have probit models for the margins.

Section 12.2: Marginal Modeling: Generalized Estimating Equations (GEE) Approach

12.2 GEE: Fitzmaurice et al. (1993), Liang et al. (1992), Molenberghs and Verbeke (2005, Chap. 8–10), and Sutradhar (2003) discussed GEE methods for categorical (primarily binary) responses. For multinomial responses, see Heagerty and Zeger (1996), Lipsitz et al. (1994), Miller et al. (1993), Parsons et al. (2009), Touloumis (2011), and references in Agresti and Natarajan (2001). More general models with ordinal responses allow for dispersion parameters that also depend on covariates (Toledano and Gatsonis 1996). LaVange et al. (2001) used GEE methods to account for clustered sampling in surveys and clinical trials.

12.3 WLS: Koch et al. (1977) used weighted least squares (WLS) to fit marginal models to [Table 12.1](#). WLS for categorical modeling is described in Section 16.7.1. It has severe limitations (e.g., covariates must be categorical and marginal tables cannot be sparse) but led naturally to the GEE approach (Miller et al. 1993).

Section 12.3: Quasi-likelihood and Its GEE Multivariate Extension: Details

12.4 Quasi-likelihood and model misspecification: Firth (1993b) gave an overview of quasi-likelihood methods. Besides McCullagh (1983), Heyde (1997) and Liang and Zeger (1995) discussed unbiased estimating functions and their connections with asymptotic consistency and efficiency. Godambe showed in 1960 that ML estimators are optimal solutions with an unbiased estimating function. When quasi-likelihood estimators are not ML, Cox (1983) and Firth (1987) suggested that they still retain good efficiency when the departure from the natural exponential family is at most moderate, such as modest overdispersion relative to such a family. The GEE methods proposed by Liang and Zeger (1986, 1995) also built on related theory in the econometrics literature about model misspecification. See Gourieroux et al. (1984), Hansen (1982), and White (1982).

12.5 GEE/ML/GEE2: The generalized estimating equations are likelihood equations, and hence the GEE estimates are also ML, in certain cases. Examples are multivariate normal data or binary data when the working covariance is correct (Fitzmaurice et al. 1993). A GEE2 analysis adds estimating equations for the correlation structure (Prentice and Zhao 1991). This has the potential to increase efficiency. A disadvantage is that, unlike with ordinary GEE, β is no longer consistent if this part of the model is misspecified.

12.6 Missing data: Surveys of ways to handle missing data include Fleiss et al. (2003, Chap. 16), Little (2005), Little and Rubin (2002), and Molenberghs and Verbeke (2005, Chap. 26–32). See also Altham (2010), Baker and Laird (1988), Fitzmaurice et al. (1994), Forster and Smith (1998), Ibrahim et al. (2005), Molenberghs and Goetghebeur (1997), Park and Brown (1994),

and Rubin (1996). Stokes et al. (2012) showed how to build the missingness pattern into a model to check whether it is associated with the response or interacts with effects of explanatory variables.

Section 12.4: Transitional Models: Markov Chain and Time Series Models

12.7 Markov chains: For statistical inference with Markov chains, see Andersen (1980, Sec. 7.7), Anderson and Goodman (1957), Billingsley (1961), Bishop et al. (1975, Chap. 7), and Kalbfleisch and Lawless (1985). Conaway (1989), Hoeting et al. (2000), Stiratelli et al. (1984), and Ware et al. (1988) proposed other analyses focusing on the conditional dependence structure.

12.8 Time series: For time series modeling of a categorical response, see Azzalini (1994), Bonney (1987), Cox (1970), Fahrmeir and Kaufmann (1987), Fokianos and Kedem (2002, 2003), Heagerty (2002), Kalbfleisch and Lawless (1985), Klingenberg (2008), Liang and Zeger (1989), Muenz and Rubinstein (1985), Stoffer et al. (1993), Varin and Vidoni (2006), Zeger and Qaqish (1988), Zhao and Prentice (1990), and the many references in Fokianos and Kedem (2003) and Klingenberg (2008). Transitional models can also incorporate latent variables (e.g., Lin et al. 2008), as discussed in Section 14.1.5.

EXERCISES

Applications

12.1 For the attitudes about abortion data of [Table 11.13](#) in Section 11.7.4, consider the model

$$\text{logit}[P(Y_t = 1)] = \alpha + \beta_1 I(t = 1) + \beta_2 I(t = 2),$$

where the indicators refer to (1) low income and (2) unmarried, the effects contrasting each with endangered woman's health.

- a. Find the ML estimates of β_1 and β_2 and their SE values by treating the three observations for a subject as independent.
- b. Find the GEE estimates and empirical SE values based on working correlations structures (i) exchangeable, (ii) independence. Why are the SE values smaller with the GEE analyses than in (a)? When would you expect them to be larger?
- c. Find the ML estimates of β_1 and β_2 and their SE values by treating the three observations for a subject as dependent. Compare to results in (a).

12.2 For [Table 10.1](#), fit a marginal model by ML or GEE to describe main effects of race, gender, and substance type (marijuana, alcohol, cigarettes) on whether a subject had used that substance. Summarize effects.

12.3 Refer to Exercise 12.2. Further study shows evidence of an interaction between gender and substance type. Using GEE with exchangeable working correlation, the model fit for the probability π of using a particular substance is

$$\text{logit}(\hat{\pi}) = -0.57 + 0.38r - 0.20g + 1.93s_1 + 0.86s_2 + 0.37g \times s_1 + 0.22g \times s_2,$$

where r , g , s_1 , s_2 are indicator variables for race (1 = white), gender (1 = female), and substance type ($s_1 = 1$, $s_2 = 0$ for alcohol; $s_1 = 0$, $s_2 = 1$ for cigarettes; $s_1 = s_2 = 0$ for marijuana). Show that:

- a. The estimated odds that a nonwhite male has used marijuana are 0.57.
- b. Given gender, the estimated odds that a white subject used a given substance are 1.46 times the estimated odds for a black subject.
- c. Given race, the estimated odds that a female has used alcohol are 1.19 times the estimated odds for males; for cigarettes and for marijuana, the estimated odds ratios are 1.02 and 0.82.
- d. Given race, the estimated odds that a female has used alcohol (cigarettes) are 9.97 (2.94) times the estimated odds she has used marijuana, and the estimated odds that a male has used alcohol (cigarettes) are 6.89 (2.36) times the estimated odds he has used marijuana. Interpret the interaction.

12.4 For [Table 12.1](#) from the depression study, fit by ML or GEE the marginal logistic model allowing treatment \times time interaction, using the time scores (1, 2, 4) for the week number. Interpret estimates. Compare substantive results to those in Section 12.1.1 for scores (0, 1, 2).

12.5 Analyze [Table 12.8](#) using a marginal logistic model with age and maternal smoking as predictors. Compare interpretations to the Markov model of Section 12.4.4.

12.6 [Table 12.9](#) refers to a three-period crossover trial to compare placebo (treatment A) with a low-dose analgesic (treatment B) and high-dose analgesic (treatment C) for relief of primary dysmenorrhea. Subjects in the study were divided randomly into six groups, the possible sequences for administering the treatments. At the end of each period, each subject rated the treatment as giving no relief (0) or some relief (1). Let $y_{i(k)t} = 1$ denote relief for subject i nested in treatment sequence k , using treatment t ($t = A, B, C$). Assuming common treatment effects for each sequence, and setting $\beta_A = 0$, obtain and interpret $\{\beta_t\}$ (using ML or GEE) for the model

[Table 12.9](#) Crossover Trial Data for Exercise 12.6

Treatment	Response Pattern for Treatments (A, B, C)							
Sequence	000	001	010	011	100	101	110	111
A B C	0	2	2	9	0	0	1	1
A C B	2	0	0	9	1	0	0	4
B A C	0	1	1	8	1	3	0	1
B C A	0	1	1	8	1	0	0	1
C A B	3	0	0	7	0	1	2	1
C B A	1	5	0	4	0	3	1	0

Source: Jones and Kenward (1987).

$$\text{logit}[P(Y_{i(k)t} = 1)] = \alpha_k + \beta_t.$$

How would you order the drugs, taking significance into account?

12.7 [Table 12.10](#) is from a Kansas State University survey of 262 pig farmers. For the question “What are your primary sources of veterinary information?,” the categories were (A) professional consultant, (B) veterinarian, (C) state or local extension service, (D) magazines, and (E) feed companies and reps. Farmers sampled were asked to select all relevant categories. The $2^5 \times 2 \times 4$ table shows the (yes, no) counts for each of these five sources cross-classified with the farmers’ education (whether they had at least some college education) and size of farm (number of pigs marketed annually, in thousands).

[Table 12.10](#) Veterinary Information Data for Exercise 12.7

Response on D																	
Educ	Pigs	E	A = yes								A = no						
			B = yes				B = no				B = yes				B = no		
			C = yes	C = no													
Ed	Pigs	E	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	
No	< 1	Y	1	0	0	0	0	0	0	2	1	1	2	1	1	5	3
		N	0	0	0	0	0	0	1	1	0	0	5	4	7	7	0
	1–2	Y	2	0	0	0	0	0	0	4	0	0	4	1	0	0	4
		N	0	0	0	0	0	0	0	0	0	0	5	0	3	4	0
	2–5	Y	3	0	0	0	0	0	0	3	0	0	1	2	0	1	1
		N	1	0	0	0	0	0	3	0	0	0	2	0	1	4	0
	> 5	Y	2	0	0	0	0	0	0	1	0	1	0	0	1	0	2
		N	1	0	0	2	1	0	1	6	0	1	1	0	0	6	0
	Some	< 1	Y	3	0	0	0	0	0	4	0	1	1	0	0	2	11
		N	0	0	0	0	0	0	0	4	0	1	2	4	6	14	0
		1–2	Y	0	0	0	0	0	0	2	0	0	1	0	0	1	6
		N	0	0	0	0	1	0	0	1	2	1	0	4	2	7	14
		2–5	Y	0	0	0	0	0	0	0	1	0	0	0	1	1	3
		N	1	0	0	0	0	0	0	0	0	0	5	0	4	4	0
		> 5	Y	1	0	0	0	0	0	0	0	0	1	1	0	0	2
		N	1	1	0	0	1	0	10	0	0	0	4	1	2	4	0

Source: Data courtesy of Tom Loughin.

- Explain why it is not proper to analyze the data by fitting a multinomial model to the counts in the $2 \times 4 \times 5$ contingency table cross-classifying education by size of farm by the source of veterinary information, treating source as the response variable. (This table contains 453 positive responses of sources from the 262 farmers.)
- For a farmer with education i and size of farm s , let $\pi_j(i_s)$ denote the probability of responding “yes” on source j . [Table 12.11](#) shows output for using GEE with exchangeable working correlation to estimate parameters in the model lacking an education effect,

[Table 12.11](#) Output for Veterinary Data of Exercise 12.7

Working Correlation Matrix					
	Col1	Col2	Col3	Col4	Col5
Row1	1.0000	0.0997	0.0997	0.0997	0.0997
Row2	0.0997	1.0000	0.0997	0.0997	0.0997
Row3	0.0997	0.0997	1.0000	0.0997	0.0997
Row4	0.0997	0.0997	0.0997	1.0000	0.0997
Row5	0.0997	0.0997	0.0997	0.0997	1.0000

Analysis Of GEE Parameter Estimates					
	Empirical Standard Error Estimates				
Parameter	Estimate	Std Error	Z	Pr > Z	
source 1	-4.4994	0.6457	-6.97	<.0001	
source 2	-0.8279	0.2809	-2.95	0.0032	
source 3	-0.1526	0.2744	-0.56	0.5780	
source 4	0.4875	0.2698	1.81	0.0708	
source 5	-0.0808	0.2738	-0.30	0.7680	
size*source 1	1.0812	0.1979	5.46	<.0001	
size*source 2	0.0792	0.1105	0.72	0.4738	
size*source 3	-0.1894	0.1121	-1.69	0.0912	
size*source 4	-0.2206	0.1081	-2.04	0.0412	
size*source 5	-0.2387	0.1126	-2.12	0.0341	

$$\text{logit}[\pi_j(is)] = \alpha_j + \beta_j s, \quad s = 1, 2, 3, 4.$$

Explain how to interpret the working correlation matrix. Explain why the results suggest a strong positive size of farm effect for source A and perhaps a weak negative size effect of similar magnitude for C, D, and E.

c. Constraining $\beta_3 = \beta_4 = \beta_5$, the ML estimate of the common slope is -0.184 ($SE = 0.063$).

Explain why it is advantageous to fit the marginal model simultaneously for all sources rather than separately to each. [Agresti and Liu (1999) and Loughin and Scherer (1998) discussed analyses for data of this form.]

12.8 Refer to [Table 13.3](#) in Section 13.3.2 on attitudes toward legalized abortion. For the response Y_t (1 = support legalization, 0 = oppose) for question t ($t = 1, 2, 3$) and for gender g (1 = female, 0 = male), consider the model: $\text{logit}[P(Y_t = 1)] = \alpha + \gamma_{og} + \beta_t$ with $\beta_3 = 0$.

a. A GEE analysis using unstructured working correlation gives correlation estimates 0.826 for questions 1 and 2, 0.797 for 1 and 3, and 0.832 for 2 and 3. What does this suggest about a reasonable working correlation structure? Why?

b. [Table 12.12](#) shows a GEE analysis with exchangeable working correlation. Interpret effects.

Table 12.12 Output for Exercise 12.8 on Abortion Attitudes

Working Correlation Matrix					
	Col1	Col2	Col3		
Row1	1.0000	0.8173	0.8173		
Row2	0.8173	1.0000	0.8173		
Row3	0.8173	0.8173	1.0000		

Analysis Of GEE Parameter Estimates					
	Empirical Standard Error Estimates				
Parameter	Estimate	Std Error	Z	Pr > Z	
Intercept	-0.1253	0.0676	-1.85	0.0637	
question 1	0.1493	0.0297	5.02	<.0001	
question 2	0.0520	0.0270	1.92	0.0544	
question 3	0.0000	0.0000	.	.	
female	0.0034	0.0878	0.04	0.9688	

c. Treating the three responses for each subject as independent observations and performing ordinary logistic regression, $\hat{\beta}_1 = 0.149$ ($SE = 0.066$), $\hat{\beta}_2 = 0.052$ ($SE = 0.066$), and $\hat{\beta}_3 = 0.004$ ($SE = 0.054$). Give a heuristic explanation of why within-subject standard errors are much larger than with GEE, yet the between-subject standard error is smaller. (See also Exercise 12.21.)

12.9 For the air pollution data in [Table 12.13](#), using ML or GEE, fit marginal logistic models that assume (a) marginal homogeneity, (b) a linear effect of time, and (c) no pattern. Interpret and compare.

Table 12.13 Results of Breath Test at Four Ages

Y_9	Y_{10}	Y_{11}	Y_{12}	Count	Y_9	Y_{10}	Y_{11}	Y_{12}	Count
1	1	1	1	94	0	1	1	1	19
1	1	1	0	30	0	1	1	0	15
1	1	0	1	15	0	1	0	1	10
1	1	0	0	28	0	1	0	0	44
1	0	1	1	14	0	0	1	1	17
1	0	1	0	9	0	0	1	0	42
1	0	0	1	12	0	0	0	1	35
1	0	0	0	63	0	0	0	0	572

Source: Ware et al. (1988).

12.10 Use GEE methods to analyze the clinical trials data in [Table 13.7](#), treating observations within each center as a correlated cluster.

12.11 For [Table 11.6](#), use GEE methods with cumulative logits to compare the two marginal distributions. Compare results to those using ML in Section 11.3.4.

12.12 For the Presidential voting summarized in [Table 11.1](#), suppose 100 men in the sample voted in the 2004 election but not in the 2008 election. Of them, in 2004, 50 voted Democrat and 50 voted Republican. Show how you can use all 533 observations to estimate the distribution of Y_1 , the 433 complete observations to estimate the distribution of $(Y_2|Y_1)$, and use these results to estimate the joint distribution. What assumption does this analysis make? Explain why the implied estimate of the odds ratio is the same as using only the complete observations.

12.13 Refer to transitional models for the insomnia study of [Table 12.3](#).

a. To compare effects while adjusting for the initial response, fit model [\(12.10\)](#), using scores $\{10, 25, 45, 75\}$ for time to falling asleep. Also fit the interaction model, and describe the lack of fit. (Note that for the first two baseline levels, the active and placebo treatments have similar sample response distributions at the follow-up; at higher baseline levels, the active treatment seems more successful.)

b. Fit the interaction model

$$\text{logit}[P(Y_2 \leq j)] = \alpha_j + \beta_1 x + \beta_2 y_1 + \beta_3(x \times y_1)$$

that constrains effects $\{\beta_1 x + \beta_2 y_1 + \beta_3 xy_1\}$ to follow the pattern $(\tau, \tau, \lambda + \sigma, \lambda)$ for the active group and $(\tau, \tau, \sigma, 0)$ for the placebo group. Interpret λ .

12.14 [Table 12.13](#) refers to a longitudinal study at Harvard of effects of air pollution on respiratory illness in children. The children were examined annually at ages 9 through 12 and classified according to the presence or absence of wheeze. Denote the binary response ($1 =$ wheeze, $0 =$ no wheeze) by Y_t at age t , $t = 9, 10, 11, 12$.

a. Explain why the loglinear model $(Y_9 Y_{10}, Y_{10} Y_{11}, Y_{11} Y_{12})$ represents a first-order Markov chain. Show that it has $G^2 = 122.9$ ($df = 8$).

b. Explain why the model $(Y_9 Y_{10} Y_{11}, Y_{10} Y_{11} Y_{12})$ represents a second-order Markov chain and satisfies conditional independence at ages 9 and 12, given states at ages 10 and 11. Show it has $G^2 = 23.9$ ($df = 4$).

c. Show that the loglinear model $(Y_9 Y_{10}, Y_9 Y_{11}, Y_9 Y_{12}, Y_{10} Y_{11}, Y_{10} Y_{12}, Y_{11} Y_{12})$ has $G^2 = 1.5$ ($df = 5$). Show that the association seems similar for pairs of ages 1 year apart, and somewhat weaker for pairs of ages more than 1 year apart. Show that the simpler model in which

$$\lambda_{ij}^{Y_9 Y_{10}} = \lambda_{ij}^{Y_{10} Y_{11}} = \lambda_{ij}^{Y_{11} Y_{12}} \quad \text{and} \quad \lambda_{ij}^{Y_9 Y_{11}} = \lambda_{ij}^{Y_9 Y_{12}} = \lambda_{ij}^{Y_{10} Y_{12}}$$

fits well, with $G^2 = 2.3$ ($df = 9$) and estimated log odds ratios of 1.75 in the first case and 1.04 in the second.

12.15 Refer to [Table 12.8](#) on respiratory illness and maternal smoking.

a. Does a transitional model with two previous responses fit better than the first-order model of Section 12.4.4? Interpret.

b. Combine the data for the two levels of maternal smoking. Does a first-order Markov chain model these data adequately? Find a loglinear model that fits adequately.

12.16 Analyze the depression data of [Table 12.1](#) using a first-order transitional model. Compare interpretations to those using marginal models.

12.17 [Table 12.14](#) is from a longitudinal study of coronary risk factors in schoolchildren (Woolson and Clarke 1984). A sample of children aged 11–13 in 1977 were classified by gender and by relative weight (obese, not obese) in 1977, 1979, and 1981. Analyze these data.

Table 12.14 Data for Exercise 12.17 on Weight Trajectories

Gender	Responses ^a							
	NNN	NNO	NON	NOO	ONN	ONO	OON	OOO
Male	119	7	8	3	13	4	11	16
Female	129	8	7	9	6	2	7	14

^aNNN indicates not obese in 1977, 1979, and 1981; NNO indicates not obese in 1977 and 1979 but obese in 1981; and so on.

Source: Reproduced with permission from the Royal Statistical Society, London (Woolson and Clarke 1984).

12.18 Analyze the pig farmer data of Exercise 12.7 using marginal models with all the variables.

12.19 Lesaffre and Spiessens (2001) analyzed data from a dermatology clinical trial for toenail infection. Analyze the data for the binary endpoint, which are available at www.blackwellpublishing.com/rss, using methods of this chapter. Prepare a two-page report summarizing your analyses, and include an appendix with relevant software output.

12.20 Results of the annual boat race between crews from Oxford and Cambridge are shown at www.theboatrace.org. Consider the time series consisting of the sequence of winners, excluding the dead heat in 1877. Analyze these data. See Section 13.3.10 for an analysis that accounts for weight differences between the crews.

Theory and Methods

12.21 Recall that positive correlation results in reduced SE values for within-cluster estimated effects. How about between-cluster effects? Suppose y_{11}, \dots, y_{1T} are Bernoulli trials with $E(y_{1j}) = \pi_1$, suppose y_{21}, \dots, y_{2T} are Bernoulli trials with $E(y_{2j}) = \pi_2$. Suppose that for $i = 1, 2$, $\text{corr}(y_{ij}, y_{ik}) = \rho$ and $\text{corr}(y_{1j}, y_{2k}) = 0$ for all $j \neq k$. Find $SE(\hat{\pi}_1 - \hat{\pi}_2)$, and show that it is larger when $\rho > 0$ than when observations within the two samples are independent.

12.22 In the example in Section 12.3.3 of a common mean model for count data, the model-based variance estimate is \bar{y}/n , whereas the sandwich estimator is $\sum_i (y_i - \bar{y})^2/n^2$. Which would you expect to be better (a) if the Poisson model holds, and (b) if there is severe overdispersion? Why?

12.23 In the example in Section 12.3.3 of a common mean model for count data, suppose we assume that $v(\mu_i) = \sigma^2$ when actually $\text{var}(Y_i) = \mu_i$. Find the model-based asymptotic variance, the actual asymptotic variance, and the sandwich estimator of the actual variance.

12.24 Consider the model of marginal homogeneity for matched-pairs binary data, expressed in form, $P(Y_t = 1) = \beta$, $t = 1, 2$. Show how the GEE expressions for the working covariance matrix [\(12.6\)](#), the estimating equations [\(12.7\)](#), and variance V_G of $\sqrt{n}\hat{\beta}$ simplify, assuming working correlation structure (a) independence, (b) exchangeable (i.e., allowing correlation, as there is only one pair).

12.25 Show that [\(12.4\)](#) is equivalent to the formula for the large-sample covariance of the ML estimator in a GLM, estimated by [\(4.31\)](#).

12.26 a. Explain the sense in which GEE methodology is a multivariate version of QL.

b. Summarize the advantages and disadvantages of the QL approach.

c. Describe conditions under which GEE parameter estimators are consistent and conditions under which they are not. For conditions in which they are consistent, explain why.

12.27 Refer to the interpretation at the end of Section 12.1.3 based on shifts in the mean for the insomnia study. State a normal latent variable model, and show how to generate a similar result by comparing estimates of the number of standard deviation shift between initial and follow-up

responses for each group.

12.28 For the analysis of the insomnia data at the end of Section 12.1.3, explain how to calculate the SE for the difference between the difference of means reported there. (Note that one difference uses paired samples and the other uses independent samples.)

12.29 For a poll of a random sample of 1800 voting-age British citizens, followed-up six months later, 794 indicated approval of the Prime Minister's performance in office each time, 570 indicated disapproval each time, 150 indicated approval initially and disapproval later, 86 indicated disapproval initially and approval later, and 200 responded at the first survey (100 approving and 100 disapproving) but not at the second survey. Denote the response by X at the first survey and by Y at the second survey. Let $\Delta_1 = P(Y=1|X=1) - P(Y=1|X=2)$ and $\Delta_2 = P(X=1|Y=1) - P(X=1|Y=2)$.

- a. Explain why the missing observations do not appear to be MCAR. [Hint: Would you expect to see this distribution of missingness if the missing observations were a random sample of all the observations?]
- b. Explain intuitively why there is insufficient information to determine whether the missing observations are MAR.
- c. Under the MAR assumption, explain intuitively why you can use the fully observed data to estimate Δ_1 and the odds ratio but not Δ_2 .
- d. Under the MAR assumption, use the information about the distribution of $(Y|X)$ from the fully observed data to predict how the missing data contribute to the cell counts. Use them together with the observed counts to estimate Δ_2 . (For details, see Fleiss et al. 2003, Sec. 16.2.)

12.30 What is wrong with this statement?: “For a first-order Markov chain, Y_t is independent of y_{t-2} .”

12.31 Gamblers A and B have a total of I dollars. They play games of pool repeatedly. In each game they each bet \$1, and the winner takes the other's dollar. The outcomes of the games are statistically independent, and A has probability π and B has probability $1 - \pi$ of winning any game. Play stops when one player has all the money. Let Y_t denote A's monetary total after t games.

- a. Show that $\{Y_t\}$ is a first-order Markov chain.
- b. State the transition probability matrix. (For this *gambler's ruin* problem, 0 and I are *absorbing* states. Eventually, the chain enters one of these. The other states are *transient*.)

12.32 For a first-order Markov chain, let X , Y , and Z denote the classifications for the $I \times I \times T$ table consisting of $\{n_{ij}(t), i = 1, \dots, I, j = 1, \dots, I, t = 1, \dots, T\}$.

- a. Explain why all transition probabilities are stationary if expected frequencies for this table satisfy loglinear model (XY, XZ) . [Thus, the likelihood-ratio statistic for testing stationary transition probabilities equals G^2 for testing fit of model (XY, XZ) .]
- b. For a Markov chain with stationary transition probabilities, let n_{ijk} denote the number of transitions from i to j to k over two successive steps. For $\{n_{ijk}\}$, argue that the goodness of fit of loglinear model $(Y_1 Y_2, Y_2 Y_3)$ tests that the chain is first-order against the alternative that it is second-order (Anderson and Goodman 1957).

CHAPTER 13

Clustered Categorical Data: Random Effects Models

In Chapter 12 we dealt with observations that occur in clusters, such as sets of repeated observations over time for subjects in a longitudinal study. Observations within clusters are usually positively correlated, tending to be more alike than observations from different clusters. Ordinary analyses that ignore the correlation and treat within-cluster observations the same as between-cluster observations produce invalid standard errors, tending to be too small for between-cluster effect estimates and too large for within-cluster effects.

In Chapter 12 we modeled the *marginal* distributions of the clustered responses, treating the joint dependence structure as a nuisance. In this chapter we present an alternative approach that adds cluster-level terms to the model that take the same value for each observation in a cluster. They are unobserved and, when treated as varying randomly among clusters, are called *random effects*. We introduced this approach in Section 11.2.4 with a model for matched pairs. The models have effects that pertain at the cluster level. We refer to such effects as *cluster-specific*, or *subject-specific* when each cluster is a subject. By contrast, in marginal models effects have *population-averaged* interpretations.

In Section 13.1 we extend the generalized linear model to include random effects, giving a *generalized linear mixed model*. In Section 13.2 we present the most important special case for binary data, the *logistic-normal model*, which uses the logit link and assumes a normal distribution for the random effects. We show several examples in Section 13.3. Section 13.4 extends this model to multinomial responses, and Section 13.5 introduces models with a hierarchical, “multilevel” structure. In Section 13.6 we discuss model fitting and in Section 13.7 the Bayesian approach to modeling multivariate categorical data.

13.1 RANDOM EFFECTS MODELING OF CLUSTERED CATEGORICAL DATA

Parameters that describe a factor's effects in an ordinary generalized linear model (GLM) are called *fixed effects*. They apply to *all* categories of interest, such as genders, treatments, or age groupings. By contrast, random effects usually apply to a *sample*. For a study that makes repeated observations on a sample of subjects, for example, the model treats observations from a given subject as a cluster, and it has a random effect for each subject.

GLMs extend ordinary regression by allowing nonnormal responses and a link function of the mean. The *generalized linear mixed model* (GLMM) is a further extension that permits random effects as well as fixed effects in the linear predictor.

13.1.1 Generalized Linear Mixed Model

Let y_{it} denote observation t in cluster i , $t = 1, \dots, T_i$. As in marginal models, the number of observations may vary by cluster. In a longitudinal study, even if clusters have equal size, many of them may have missing observations. Let \mathbf{x}_{it} denote a column vector of values of explanatory variables for this observation.

Let \mathbf{u}_i denote the vector of random effects for cluster i . Often, the random effect is univariate. Conditional on \mathbf{u}_i , a GLMM resembles an ordinary GLM. Let $\mu_{it} = E(Y_{it}|\mathbf{u}_i)$. The linear predictor for a GLMM has the form

$$(13.1) \quad g(\mu_{it}) = \mathbf{x}_{it}^T \boldsymbol{\beta} + z_{it}^T \mathbf{u}_i$$

for link function $g(\cdot)$ and fixed effect model parameters $\boldsymbol{\beta}$. The random effect vector \mathbf{u}_i is assumed to have a multivariate normal distribution $N(0, \Sigma)$. The covariance matrix Σ depends on unknown *variance components* and possibly also correlation parameters. Conditional on \mathbf{u}_i , standard model fitting treats $\{y_{it}\}$ as independent over i and t . As discussed in Section 11.2.2, the variability among \mathbf{u}_i induces nonnegative associations among the responses, for the marginal distribution averaged over the subjects. This is caused by the shared random effect \mathbf{u}_i for each observation in a cluster.

In (13.1), the random effect enters the model on the same scale as the predictor terms. This is convenient but also natural for many applications. For instance, random effects sometimes represent heterogeneity caused by omitting certain explanatory variables. Consider the special case with univariate random effect and $z_{it} = 1$. With u_i replaced by $u_i^* \sigma$ where $\{u_i^*\}$ are $N(0, 1)$, the GLMM has the form

$$g(\mu_{it}) = \mathbf{x}_{it}^T \boldsymbol{\beta} + u_i^* \sigma.$$

This has the form of an ordinary GLM with unobserved values $\{u_i^*\}$ of a particular covariate. Thus, random effects models relate to methods of dealing with unmeasured predictors and other forms of missing data. The random effects part of the linear predictor reflects terms that would be in the fixed effects part if those explanatory variables had been included. Random effects also sometimes represent random measurement error in the explanatory variables. If we replace a particular x_{it} by $x_{it}^* + \varepsilon_i$, with x_{it}^* the true value and ε_i the measurement error, then ε_i times the regression parameter can be absorbed in the random effects term. Related to these motivations, random effects also provide a mechanism for explaining overdispersion in basic models not having those effects (Breslow and Clayton 1993, Molenberghs et al. 2010).

13.1.2 Logistic GLMM with Random Intercept for Binary Matched Pairs

We illustrate the GLMM expression (13.1) using a simple case, that of binary matched pairs. Cluster i consists of the responses (y_{i1}, y_{i2}) for matched pair i . Observation t in cluster i has $y_{it} = 1$ (a success) or 0 (a failure), $t = 1, 2$.

In Section 11.2.2 we introduced the model

$$(13.2) \text{logit}[P(Y_{it} = 1)] = \alpha_i + \beta x_t,$$

where $x_1 = 0$ and $x_2 = 1$. For it, β is a cluster-specific log odds ratio. That section treated α_i as a fixed effect and eliminated it using conditional ML. An equivalent representation of (13.2) is

$$(13.3) \text{logit}[P(Y_{i1} = 1|u_i)] = \alpha + u_i, \quad \text{logit}[P(Y_{i2} = 1|u_i)] = \alpha + \beta + u_i,$$

where $u_i = \alpha_i - \alpha$ for some constant α . Now, we treat u_i as a random effect for cluster i , with $\{u_i\}$ independent from a $N(0, \sigma^2)$ distribution with σ unknown. Conditionally on u_i , we treat y_{i1} and y_{i2} as independent.

Model (13.3) is the special case of (13.1) in which $\mu_{it} = P(Y_{it} = 1|u_i)$, $g(\cdot)$ is the logit link, $\beta^T = (\alpha, \beta)$, $x_{i1}^T = (1, 0)$ and $x_{i2}^T = (1, 1)$ for all i , and $z_{it} = 1$ for all i and t . The univariate random effect adjusts the intercept but does not modify the fixed effect. A GLMM with random effect of this form is called a *random intercept model*. Instead of the usual fixed intercept α , it has a random intercept $\alpha + u_i$.

Let $Y_1 = \sum_i y_{i1}$ and $Y_2 = \sum_i y_{i2}$. These are dependent binomial variates. Marginally, Y_1 is binomial with n trials and parameter $E\{\exp(\alpha + U)/[1 + \exp(\alpha + U)]\}$, and Y_2 is binomial with parameter $E\{\exp(\alpha + \beta + U)/[1 + \exp(\alpha + \beta + U)]\}$. The expectations refer to U , a $N(0, \sigma^2)$ random variable. The model implies a nonnegative correlation between Y_1 and Y_2 , with greater association resulting from greater heterogeneity (i.e., larger σ). Clusters with a large positive u_i have a relatively large $P(Y_{it} = 1|u_i)$ for each t , whereas clusters with a large negative u_i have a relatively small $P(Y_{it} = 1|u_i)$ for each. For this model, Y_1 and Y_2 are independent only if $\sigma = 0$.

A 2×2 population-averaged table with (success, failure) for both the row and column categories summarizes the number of observations for which $(y_{i1}, y_{i2}) = (1, 1), (1, 0), (0, 1)$, or $(0, 0)$. Let $\{n_{ab}\}$ denote these counts. Table 13.1 is an example. Let $\{\hat{\mu}_{ab}\}$ denote marginal fitted values for model (13.3). We defer discussion of model fitting until Section 13.6. However, model (13.3) is a rare instance in which the fixed effect in a model containing random effects has a closed-form ML estimate,

Table 13.1 Presidential Votes in 2004 and in 2008, for Males Sampled in 2010 by the General Social Survey

		2008 Election		
2004 Election	Democrat	Republican	Total	
Democrat	175	16	191	
Republican	54	188	242	
Total	229	204	433	

$$\hat{\beta} = \log(\hat{\mu}_{21}/\hat{\mu}_{12}).$$

When the sample log odds ratio $\log(n_{11}n_{22}/n_{12}n_{21}) \geq 0$, then $\{\hat{\mu}_{ab} = n_{ab}\}$ and $\hat{\beta} = \log(n_{21}/n_{12})$. This is the same as the conditional ML estimate (Section 11.2.3). Neuhaus et al. (1994) showed that this is true for any parametric choice of random effects distribution for which the model (13.3) can generate $\{n_{ab}\}$ as fitted values. Lindsay et al. (1991) showed that this estimate also results with a nonparametric approach discussed in Section 14.2.5. The model implies that the true log odds ratio for this 2×2 table is at least 0. When $\log(n_{11}n_{22}/n_{12}n_{21}) < 0$, however, then $\hat{\sigma} = 0$ and the fitted values

$\{\hat{\mu}_{ab} = n_{a+b}/n\}$ satisfy independence. Then, $\hat{\beta}$ is identical to the estimate for the marginal model (11.6) by which β is the difference between logits for the two marginal distributions, which is the log odds ratio $\hat{\beta} = \log[(n_{+1}/n_{+2})/(n_{1+}/n_{2+})]$.

13.1.3 Example: Changes in Presidential Voting Revisited

[Table 13.1](#) on voting in successive Presidential elections was first analyzed in Section 11.1. For it, the ML fit of model [\(13.3\)](#) yields $\hat{\beta} = \log(54/16) = 1.216$ ($SE = 0.285$), with $\hat{\sigma} = 5.22$ describing variability of the random effects. This is identical to the conditional ML estimate [\(11.10\)](#), with standard error $[(1/54) + (1/16)]^{1/2} = 0.285$. For a given male voting Democrat or Republican in these two elections, the estimated odds of voting Democrat in 2008 equal $\exp(1.216) = 3.375$ times the odds in 2004. The large $\hat{\sigma}$ reflects the very strong association between the two responses, with sample odds ratio 38.1.

13.1.4 Extension: Rasch Model and Item Response Models

An extension of the logistic matched-pairs model (13.3) allows $T > 2$ observations in each cluster. The random intercept model then has form

$$(13.4) \text{ logit}[P(Y_{it} = 1|u_i)] = \alpha + \beta_t + u_i,$$

where $\{u_i\}$ are independent $N(0, \sigma^2)$ and where identifiability requires a constraint such as $\beta_1 = 0$. Equivalently, the model can delete α and the constraint on $\{\beta_t\}$.

Early applications of this GLMM were in psychometrics. The model describes responses to a battery of T questions on an exam. The probability $P(Y_{it} = 1 | u_i)$ that subject i makes the correct response on question t depends on the overall ability of subject i , characterized by u_i , and the easiness of question t , characterized by β_t . Such models are called *item-response models*. The logistic form (13.4) is called the *Rasch model* (Rasch 1961). In estimating $\{\beta_t\}$, Rasch treated $\{u_i\}$ as fixed effects and used conditional ML, as outlined in Section 11.2.3 for matched pairs. Later authors used the normal random effects approach for this model and sometimes the corresponding model with probit link (Bock and Aitkin 1981).

The $\{\beta_t\}$ in the Rasch model differ from parameters in corresponding marginal models such as (11.33), which is

$$(13.5) \text{ logit}[P(Y_t = 1)] = \alpha + \beta_t, \quad t = 1, \dots, T,$$

since the Rasch model effects are subject-specific. The Rasch model refers to a $T \times 2 \times n$ table of observation-by-outcome-by-subject, whereas the marginal model refers to the $T \times 2$ observation-by-outcome table of the T marginal distributions, collapsed over subjects. For observations s and t for a given subject i with model (13.4),

$$\beta_s - \beta_t = \text{logit}[P(Y_{is} = 1|u_i)] - \text{logit}[P(Y_{it} = 1|u_i)],$$

which is a log odds ratio conditional on the subject. By contrast, the corresponding population-averaged effect in marginal model (13.5) is

$$\beta_s - \beta_t = \text{logit}[P(Y_{hs} = 1)] - \text{logit}[P(Y_{ht} = 1)],$$

with subject h randomly selected for observation s and subject i randomly selected for observation t (i.e., h and i are *independent* observations).

13.1.5 Random Effects Versus Conditional ML Approaches

Suppose we treat $\{u_i\}$ in model (13.4) as fixed effects instead of random effects and use ordinary ML to estimate $\{\beta_t\}$ and $\{u_i\}$. As n increases, so does the number of parameters, since each cluster has a u_i . Even though the number of $\{\hat{\beta}_t\}$ does not increase as n does, the ordinary ML estimators $\{\hat{\beta}_t\}$ are not consistent. This happens in many models when the number of parameters has size on the same order as the number of observations. Asymptotic optimality properties of ML estimators, such as consistency, require the number of parameters to be fixed as n increases. For model (13.4), ML estimators of $\{\beta_t\}$ have bias of order $T/(T - 1)$ (Andersen 1980, pp. 244–245). For the matched-pairs model (13.2), for instance, $\hat{\beta} \rightarrow 2\beta$ in probability (Exercise 11.29).

For this reason, the preferable approach for the fixed effects model is *conditional ML*, eliminating $\{u_i\}$ by conditioning on their sufficient statistics $\{S_i = \sum_t y_{it}, i = 1, \dots, n\}$. In the item-response context, these are the numbers of correct responses for each subject. Conditional on $\{S_i\}$, the distribution of $\{y_{it}\}$ is independent of $\{u_i\}$. Maximizing the resulting likelihood then yields consistent estimators of $\{\beta_t\}$. The analysis generalizes the one in Section 11.2.3 for the subject-specific logistic model (11.8) for matched pairs.

Compared with the random effects approach, the conditional ML approach has certain advantages. It is not necessary to assume a parametric distribution for $\{u_i\}$. It is difficult to check this assumption in the random effects approach. Conditional ML is also appropriate with retrospective sampling. In that case, bias can occur with a random effects approach because the clusters are not randomly sampled (Neuhaus and Jewell 1990b).

However, the conditional ML approach has severe disadvantages. It is restricted to the canonical link (the logit), for which reduced sufficient statistics exist for $\{u_i\}$. Also, as discussed in Section 11.2.7, it is restricted to inference about within-cluster fixed effects. The conditioning removes the source of variability needed for estimating between-cluster effects in models with explanatory variables such as those considered next. Also, this approach does not provide information about $\{u_i\}$, such as predictions of their values and estimates of their variability or of the probabilities they determine. Finally, in more general models with covariates, conditional ML can be less efficient than the random effects approach for estimating the fixed effects (see Note 13.2).

13.2 BINARY RESPONSES: LOGISTIC-NORMAL MODEL

The item-response model (13.4) with random intercept is a special case of an important class of random effects models for binary data called *logistic-normal models*. With univariate random effect, the model form is

$$(13.6) \text{ logit}[P(Y_{it} = 1|u_i)] = \mathbf{x}_{it}^T \boldsymbol{\beta} + u_i,$$

where $\{u_i\}$ are independent $N(0, \sigma^2)$ variates. The logistic-normal model has a long history, dating at least to Cox (1970, Prob. 20 in that text) for the matched-pairs model (13.3) and Pierce and Sands (1975).

13.2.1 Shared Random Effect Implies Nonnegative Marginal Correlations

More generally, the link function in the random intercept model (13.6) can be an arbitrary inverse cdf. For such models, Y_{is} and Y_{it} are treated conditionally (given u_i) as independent. The model implies that they are marginally nonnegatively correlated. Let Φ denote the cdf that is the inverse link function. Then, for $s \neq t$,

$$\begin{aligned}\text{cov}(Y_{is}, Y_{it}) &= E[\text{cov}(Y_{is}, Y_{it}|u_i)] + \text{cov}[E(Y_{is}|u_i), E(Y_{it}|u_i)] \\ (13.7) \quad &= 0 + \text{cov}[\Phi(\mathbf{x}_{is}^T \boldsymbol{\beta} + u_i), \Phi(\mathbf{x}_{it}^T \boldsymbol{\beta} + u_i)].\end{aligned}$$

The functions in the last covariance term are both monotone increasing in u_i , and hence are nonnegatively correlated.

When the predictor value \mathbf{x}_{it} is the same for each t , the marginal distribution implied by the model is exchangeable among components of a cluster. This is often plausible. In longitudinal studies, however, observations closer together in time may tend to be more highly correlated.

Usually, the main focus in using a GLMM is inference about the fixed effects. The random effects part of the model is a mechanism for representing how the positive correlation occurs between observations within a cluster. Parameters pertaining to the random effects may themselves be of interest, however. For instance, the estimate $\hat{\sigma}$ of the standard deviation of a random intercept summarizes the degree of heterogeneity of a population.

13.2.2 Interpreting Heterogeneity in Logistic-Normal Models

When $\sigma = 0$, the logistic-normal model (13.6) simplifies to the ordinary logistic regression model treating all observations as independent. When $\sigma > 0$, how can we interpret the variability in effects that this model implies?

Consider observation y_{it} at setting \mathbf{x}_{it} of predictors and observation y_{hs} at setting \mathbf{x}_{hs} . Their log odds ratio is

$$\text{logit}[P(Y_{it} = 1|u_i)] - \text{logit}[P(Y_{hs} = 1|u_h)] = (\mathbf{x}_{it} - \mathbf{x}_{hs})^T \boldsymbol{\beta} + (u_i - u_h).$$

We cannot observe $(u_i - u_h)$, which has a $N(0, 2\sigma^2)$ distribution. However, $100(1 - \sigma)\%$ of those log odds ratios fall within

$$(13.8) \quad (\mathbf{x}_{it} - \mathbf{x}_{hs})^T \boldsymbol{\beta} \pm z_{\alpha/2} \sqrt{2} \sigma.$$

When $\sigma = 0$, $(\mathbf{x}_{it} - \mathbf{x}_{hs})^T \boldsymbol{\beta}$ is the usual form of log odds ratio for a model without random effects. When $\sigma > 0$, $(\mathbf{x}_{it} - \mathbf{x}_{hs})^T \boldsymbol{\beta}$ is the log odds ratio for two observations in the same cluster ($h = i$) or with the same random effect value. Suppose that $\mathbf{x}_{it} = \mathbf{x}_{hs}$ for observations from different clusters. Then, since $z_{0.25} = 0.674$, the middle 50% of the log odds ratios fall within $\pm 0.674\sqrt{2}\sigma = \pm 0.95\sigma$. Hence, the median odds ratio between the observation with higher random effect and the observation with lower random effect equals $\exp(0.95\sigma)$. With a single predictor and $x_{it} - x_{hs} = 1$, the median such odds ratio equals $\exp(\beta + 0.95\sigma)$. Larsen et al. (2000) presented related interpretations.

13.2.3 Connections Between Random Effects Models and Marginal Models

The fixed effects parameters β in GLMMs have conditional interpretations, given the random effect. Those fixed effects are of two types. First, consider an explanatory variable that varies in value among observations in a cluster. For instance, in a crossover study comparing T drugs, for each subject the drug taken varies from observation to observation in that subject's cluster of T observations. For such an explanatory variable, its coefficient in the model refers to the effect on the response of a within-cluster (e.g., subject-specific) 1-unit increase of that predictor. The random effect as well as other explanatory variables in the model are constant while that predictor increases by 1. The effect of that explanatory variable is a “within-cluster” (e.g., within-subject) one.

Second, consider an explanatory variable with constant value among observations in a cluster. An example is gender when each cluster is an individual. For such an explanatory variable, its coefficient refers to the effect on the response of a “between-cluster” 1-unit increase of that predictor. An example is a comparison of females and males using an indicator variable and its coefficient. However, this fixed effect in the GLMM applies only when the random effect (as well as other explanatory variables in the model) takes the same value in both groups: for instance, a male and a female with the same random effect values.

It is in this sense that random effects models are cluster-specific models, as both within- and between-cluster effects apply conditional on the random effect value. By contrast, effects in marginal models are averaged over all clusters (i.e., population averaged), so those effects do not refer to a comparison at a fixed value of a random effect. In fact, a fundamental difference between the two model types is that when the link function is nonlinear, such as the logit, the population-averaged effects of marginal models often are smaller in absolute value than the cluster-specific effects of GLMMs.

Specifically, the GLMM (13.1) refers to the conditional mean, $\mu_{it} = E(Y_{it}|\mathbf{u}_i)$. By inverting the link function,

$$E(Y_{it}|\mathbf{u}_i) = g^{-1}(\mathbf{x}_{it}^T \boldsymbol{\beta} + \mathbf{z}_{it}^T \mathbf{u}_i).$$

Marginally, averaging over the random effects, the mean is

$$E(Y_{it}) = E[E(Y_{it}|\mathbf{u}_i)] = \int g^{-1}(\mathbf{x}_{it}^T \boldsymbol{\beta} + \mathbf{z}_{it}^T \mathbf{u}_i) f(\mathbf{u}_i; \Sigma) d\mathbf{u}_i,$$

where $f(\mathbf{u}; \Sigma)$ is the $N(\mathbf{0}, \Sigma)$ density function for the random effects. For the identity link,

$$E(Y_{it}) = \int (\mathbf{x}_{it}^T \boldsymbol{\beta} + \mathbf{z}_{it}^T \mathbf{u}_i) f(\mathbf{u}_i; \Sigma) d\mathbf{u}_i = \mathbf{x}_{it}^T \boldsymbol{\beta}.$$

The marginal model has the same model form and effects $\boldsymbol{\beta}$. This is not true for other links. For instance, for the logistic-normal model (13.6),

$$E(Y_{it}) = E \left[\frac{\exp(\mathbf{x}_{it}^T \boldsymbol{\beta} + u_i)}{1 + \exp(\mathbf{x}_{it}^T \boldsymbol{\beta} + u_i)} \right].$$

This expectation does not have form $\exp(\mathbf{x}_{it}^T \boldsymbol{\beta})/[1 + \exp(\mathbf{x}_{it}^T \boldsymbol{\beta})]$ except when u_i has a degenerate distribution ($\sigma = 0$).

Approximate relationships exist between effects from the two model types. In the logistic-normal case with effect $\boldsymbol{\beta}$ and small σ , Zeger et al. (1988) showed that

$$(13.9) \quad E(Y_{it}) \approx \exp(c \mathbf{x}_{it}^T \boldsymbol{\beta}) / [1 + \exp(c \mathbf{x}_{it}^T \boldsymbol{\beta})],$$

where $c = [1 + 0.346\sigma^2]^{-1/2}$. Since the effect in the marginal model multiplies that of the random effects model by about c , it is smaller in absolute value. The discrepancy increases as σ increases. For $\boldsymbol{\beta}$ near 0, Neuhaus et al. (1991) showed that the marginal model effect is approximately $\boldsymbol{\beta}(1 - \rho)$, where $\rho = \text{corr}(Y_{it}, Y_{is})$ at $\boldsymbol{\beta} = \mathbf{0}$. Again, the discrepancy increases as σ increases, since ρ increases with σ .

For Table 13.1 on voting in 2004 and in 2008, the ML estimate for model (13.3) is $\hat{\boldsymbol{\beta}} = 1.216$, with

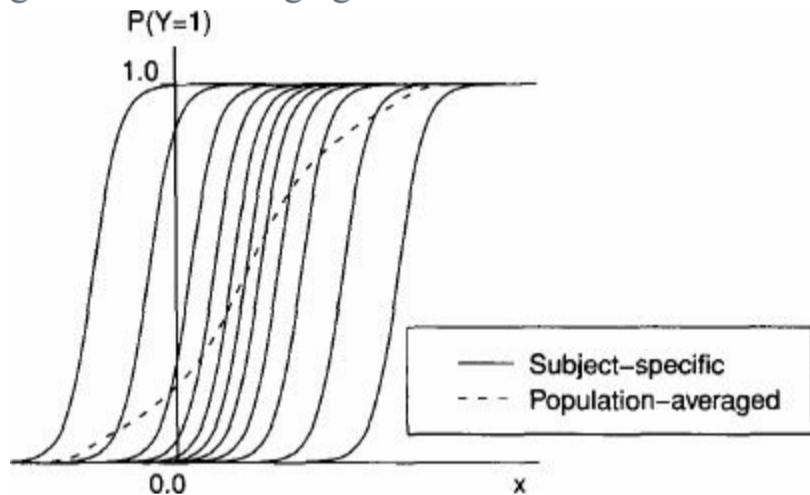
$\hat{\sigma} = 5.22$ for variability of $\{u_i\}$. Approximation (13.9) suggests that this corresponds to a marginal estimate of about $[1 + 0.346(5.22)^2]^{-1/2}(1.216) = 0.377$. The actual marginal estimate is the log odds ratio for the sample marginal distributions, equaling

$$\log[(229/204)/(191/242)] = 0.35.$$

In fact, the marginal effect (0.35) is much smaller than the subject-specific effect (1.216). At $\beta = 0$, the fit of the model is that of the symmetry model, for which $\hat{\mu}_{12} = \hat{\mu}_{21} = (n_{12} + n_{21})/2$. The correlation for that 2×2 table equals 0.676, from which the subject-specific estimate of 1.216 and the Neuhaus et al. (1991) approximation suggests a marginal estimate of about $1.216(1 - 0.676) = 0.39$, also not far from the actual value of 0.35.

[Figure 13.1](#) illustrates why the marginal effect is smaller than the subject-specific effect. For a single explanatory variable x , the figure shows subject-specific curves for $P(Y_{it} = 1|u_i)$ for several subjects when considerable heterogeneity exists. This corresponds to a relatively large σ for random effects. At any fixed value of x , variability occurs in the conditional means, $E(Y_{it}|u_i) = P(Y_{it} = 1|u_i)$. The average of these is the marginal mean, $E(Y_{it})$. These averages for various x values yield the superimposed dashed curve. It has a shallower slope. In fact, it does not exactly follow the logistic formula.

Figure 13.1 Logistic random-intercept model, showing its subject-specific curves and the population-averaged marginal curve averaging over these.



Similar remarks apply to other GLMMs. For the probit link with binary data, however, the probit model with normal random effect does imply a marginal model of probit form (Exercise 13.25). With univariate random intercept, the marginal effect equals the subject-specific effect multiplied by $[1 + \sigma^2]^{-1/2}$ (Zeger et al. 1988, Caffo and Griswold 2006). The marginal model is also of probit form when the random effects have a mixture of normal distributions (Caffo et al. 2007).

13.2.4 Comments About GLMMs Versus Marginal Models

GLMMs with random effects describe cluster-specific effects, whereas marginal models describe population-averaged effects. Some statisticians prefer one of these types, but most feel that both are useful, depending on the application.

The random effects modeling approach is preferable if we want to specify a mechanism that could generate positive association among clustered observations, estimate cluster-specific effects, estimate their variability, or model the joint distribution. Latent variable constructions used to motivate model forms (e.g., for binary data, the tolerance model and the related threshold and utility models of Section 7.1.1) apply more naturally at the cluster level than at the marginal level. Given a random effects model, we can recover information about marginal distributions. That is, a random effects model implies a marginal model, but a marginal model does not itself imply¹ a random effects model.

In many surveys or epidemiological studies, a goal is to compare the relative frequency of occurrence of some outcome for different groups in a population, such as smokers and nonsmokers. Then, quantities of primary interest include between-group odds ratios among marginal probabilities for the different groups. That is, effects of interest are between-cluster rather than within-cluster. When marginal effects are the main focus, it is simpler and often preferable to model the margins directly. We can then parameterize the model so that regression parameters have a direct marginal interpretation. Developing a more detailed model of the joint distribution that generates those margins, as a random effects model does, provides greater opportunity for misspecification. For instance, the assumptions that observations in a cluster are independent, given the random effect, and that the random effects have constant variance throughout the space of explanatory variable values, need not be valid.

In Section 13.2.3 we noted that cluster-specific effects are usually larger in magnitude than marginal effects, and the discrepancy increases as variance components increase. Usually, though, the significance of an effect is similar in the two model types. If one effect seems more important than another in a random effects model, the same is usually true with a marginal model. So the choice of the model is usually not crucial to inferential conclusions.

This statement requires a caveat, however, since sizes of effects in marginal models depend on the degree of heterogeneity in random effects models. In comparing effects for two groups or two variables that have quite different variance components, relative sizes of effects will differ for marginal and random effects models. Section 13.3.8 shows an example. From approximation (13.9), the attenuation from the random effects to the marginal effect will tend to be greater for the group having the larger variance component. For instance, suppose that two groups, one young in age and the other elderly, both show the same subject-specific effect in a crossover study comparing two drugs. If the elderly group has more heterogeneity in their response propensities, their marginal effect may be smaller than that for the younger group. The marginal effects differ even though the subject-specific effects are the same, because of the greater variance component for the elderly. In such cases, the subject-specific effect (appropriately modeled) may have more relevance.

13.3 EXAMPLES OF RANDOM EFFECTS MODELS FOR BINARY DATA

In the next three sections we present a variety of examples of random effects models. In this section we present models for binary responses.

13.3.1 Example: Small-Area Estimation of Binomial Proportions

Small-area estimation refers to estimation of parameters for a large number of geographical areas when each has relatively few observations. Examples are county-specific estimates of characteristics such as the proportions of people unemployed, living below the poverty level, and not having health insurance coverage. With a national or statewide survey, some counties may have few observations. Then, sample proportions in the counties may poorly estimate the true countywide proportions. Random effects models that treat each county as a cluster can provide improved estimates. In assuming that the true proportions vary according to some distribution, the fitting process “borrows from the whole”—it uses data from all the counties to estimate the proportion in any given one.

Let π_i denote the true proportion in area i , $i = 1, \dots, n$. These areas may be all the ones of interest, or only a sample. Let $\{y_i\}$ denote independent $\text{bin}(T_i, \pi_i)$ variates; that is, $y_i = \sum_{t=1}^{T_i} y_{it}$, where $\{y_{it}\}$, $t = 1, \dots, T_i$ are independent with $P(Y_{it} = 1) = \pi_i$ and $P(Y_{it} = 0) = 1 - \pi_i$. The sample proportions $\{p_i = y_i/T_i\}$ are ML estimates of $\{\pi_i\}$ for the fixed effects model

$$\text{logit}(\pi_i) = \alpha + \beta_i, \quad i = 1, \dots, n.$$

This model is saturated, having n nonredundant parameters (with a constraint such as $\beta_1 = 0$) for the n binomial observations.

For small $\{T_i\}$, $\{p_i\}$ have large standard errors. Thus, $\{p_i\}$ may display much more variability than $\{\pi_i\}$, especially when $\{\pi_i\}$ are similar. Then, it is helpful to shrink $\{p_i\}$ toward their overall mean. We can accomplish this with the random effects model

$$(13.10) \quad \text{logit}[P(Y_{it} = 1|u_i)] = \alpha + u_i,$$

where $\{u_i\}$ are independent $N(0, \sigma^2)$ variates. This model is a logistic analog of one-way random effects ANOVA. When $\sigma = 0$, all π_i are identical. For this model,

$$\hat{\pi}_i = \exp(\hat{\alpha} + \hat{u}_i) / [1 + \exp(\hat{\alpha} + \hat{u}_i)].$$

The predicted random effect \hat{u}_i is the estimated mean of the distribution of u_i , given the data (Section 13.6.6). This prediction depends on all the data, not just data from area i . A benefit is potential reduction in the mean squared error of the estimates around $\{\pi_i\}$, using $\{\hat{\pi}_i\}$ instead of $\{p_i\}$.

The random effects model estimate $\hat{\pi}_i$ can differ substantially from the sample proportion p_i . For example, if $\sigma = 0$, then all $\hat{u}_i = 0$. The random effects model estimate is then $\hat{\pi}_i = (\sum_{i=1}^n \sum_{t=1}^{T_i} y_{it}) / (\sum_i T_i)$, which is the overall sample proportion after pooling all n samples. When truly all π_i are equal, this is a much better estimator of that common value than the sample proportion from a single sample. Generally, the random effects model estimators shrink the separate sample proportions toward the overall sample proportion. The amount of shrinkage decreases as σ increases. The shrinkage also decreases as the $\{T_i\}$ grow. As each sample has more data, we put more trust in the separate sample proportions.

We illustrate model (13.10) with a simulated sample of 2000 people to mimic a poll taken before the 2008 U.S. presidential election. For T_i observations in state i ($i = 1, \dots, 51$, where $i = 51$ is District of Columbia = DC), y_i is $\text{bin}(T_i, \pi_i)$, where π_i is the actual proportion of votes in state i for Barack Obama in the 2008 election. Here, we took T_i proportional to the number of people in state i who voted in that election, subject to $\sum_i T_i = 2000$. Table 13.2 shows $\{T_i\}$, $\{\pi_i\}$, and $\{p_i = y_i/T_i\}$.

Table 13.2 Estimates^a of Proportion of Vote for Obama in 2008 U.S. Presidential Election, Based on Sample Size T_i in State i

State	T_i	π_i	p_i	$\hat{\pi}_i$	State	T_i	π_i	p_i	$\hat{\pi}_i$
AK	5	0.379	0.600	0.524	MT	7	0.471	0.429	0.489
AL	29	0.387	0.310	0.398	NC	66	0.497	0.348	0.390
AR	17	0.389	0.118	0.344	ND	5	0.445	0.400	0.488
AZ	35	0.449	0.371	0.425	NE	12	0.416	0.833	0.618
CA	207	0.609	0.623	0.612	NH	11	0.541	0.273	0.431
CO	37	0.537	0.432	0.461	NJ	59	0.571	0.542	0.533
CT	25	0.606	0.560	0.535	NM	13	0.569	0.538	0.519
DC	4	0.925	1.000	0.580	NV	13	0.552	0.467	0.491
DE	6	0.619	0.667	0.541	NY	116	0.629	0.664	0.638
FL	128	0.509	0.570	0.561	OH	87	0.514	0.552	0.543
GA	60	0.469	0.450	0.466	OK	22	0.344	0.182	0.350
HI	7	0.718	0.857	0.589	OR	28	0.568	0.500	0.503
IA	23	0.539	0.391	0.449	PA	92	0.545	0.576	0.562
ID	10	0.359	0.100	0.385	RI	7	0.629	0.857	0.589
IL	84	0.618	0.536	0.530	SC	29	0.449	0.310	0.398
IN	42	0.498	0.476	0.487	SD	6	0.448	0.667	0.541
KS	19	0.415	0.421	0.468	TN	40	0.418	0.425	0.455
KY	28	0.412	0.357	0.425	TX	123	0.436	0.496	0.498
LA	30	0.399	0.367	0.428	UT	15	0.342	0.667	0.570
MA	47	0.618	0.426	0.452	VA	57	0.526	0.491	0.496
MD	40	0.619	0.725	0.645	VT	5	0.675	0.800	0.560
ME	11	0.577	0.818	0.607	WA	46	0.573	0.674	0.618
MI	76	0.573	0.553	0.542	WI	45	0.562	0.578	0.554
MN	44	0.541	0.500	0.503	WV	11	0.425	0.545	0.520
MO	45	0.492	0.556	0.539	WY	4	0.325	0.500	0.506
MS	20	0.430	0.550	0.527					

^a π_i , True; p_i , sample; $\hat{\pi}_i$, estimate using random effects model.

For the ML fit of model (13.10), $\hat{\alpha} = 0.030$ and $\hat{\sigma} = 0.419$. The predicted random effect values yield the proportion estimates $\{\hat{\pi}_i\}$, also shown in Table 13.2. Since $\{T_i\}$ are mostly small and since $\hat{\sigma}$ is relatively small, considerable shrinkage of these estimates occurs from the sample proportions toward the overall proportion of Obama voters for these 2000 people sampled, which was 0.5245. The $\{\hat{\pi}_i\}$ vary between 0.344 (for Arkansas) and 0.645 (for Maryland), whereas the sample proportions vary between 0.100 (for Idaho) and 1.0 (for DC). Sample proportions based on fewer observations, such as DC, tended to shrink more. The random effects model estimates tend to be closer than the sample proportions to the true values. The root mean square error about the true proportions is 0.038 for the model-based estimates and 0.069 for the sample proportions.

13.3.2 Modeling Repeated Binary Responses: Attitudes About Abortion

In Section 13.1.4 we introduced a random effects version of the Rasch model for repeated binary measurement. This model extends to incorporate covariates.

We illustrate using [Table 13.3](#). The subjects indicated whether they supported legalizing abortion in each of three situations. [Table 13.3](#) also classifies the subjects by gender. Let y_{it} denote the response for subject i in situation t , with $y_{it} = 1$ representing support. Consider the model

Table 13.3 Support for Legalizing Abortion in Three Situations, by Gender

Gender	Sequence of Responses in Three Situations ^a							
	(1,1,1)	(1,1,0)	(0,1,1)	(0,1,0)	(1,0,1)	(1,0,0)	(0,0,1)	(0,0,0)
Male	342	26	6	21	11	32	19	356
Female	440	25	14	18	14	47	22	457

^aSituations are (1) if the family has a very low income and cannot afford any more children, (2) when the woman is not married and does not want to marry the man, and (3) when the woman wants it for any reason. 1, yes; 0, no.
Source: Data from General Social Survey.

$$(13.11) \text{logit}[P(Y_{it} = 1|u_i)] = \alpha + \beta_t + \gamma x_i + u_i,$$

where $x_i = 1$ for females and 0 for males, $\{u_i\}$ are independent $N(0, \sigma^2)$, and $\{\beta_t\}$ satisfy a constraint such as $\beta_1 = 0$. Here, the gender effect γ is assumed to be identical for each situation, and $\{\beta_t\}$ refer to the situations.

Since model (13.11) implies nonnegative association among responses in the various situations, we should use questions and scales for which this happens. With scale (yes, no), it would not be sensible for one question to ask “Should abortion be legal when a woman is not married?” and another to ask “Should abortion be illegal during the last three months of pregnancy?”

[Table 13.4](#) summarizes ML fitting results. The contrasts of $\{\beta_t\}$ indicate greater support for legalized abortion in situation 1 (when the family has low income) than in the other two. There is slight evidence of greater support in situation 2 than in situation 3. The fixed effects estimates have log odds ratio interpretations, within-subject for situation effects and between-subject for the gender effect. For a given subject of either gender, for instance, the estimated odds of supporting legalized abortion in situation 1 equal $\exp(0.835) = 2.30$ times the estimated odds in situation 3. Since $\gamma = 0.013$, for each situation the estimated probability of supporting legalized abortion is similar for females and males having the same random effect values.

Table 13.4 Summary of ML Estimates for Generalized Linear Mixed Model (13.11) for [Table 13.3](#), and ML and GEE Estimates for Corresponding Marginal Model

Effect	Parameter	GLMM ML		Marginal ML		Marginal GEE	
		Estimate	SE	Estimate	SE	Estimate	SE
Abortion	$\beta_1 - \beta_3$	0.835	0.160	0.148	0.030	0.149	0.030
	$\beta_1 - \beta_2$	0.542	0.157	0.098	0.027	0.097	0.028
	$\beta_2 - \beta_3$	0.292	0.157	0.049	0.027	0.052	0.027
Gender	γ	0.013	0.490	0.005	0.088	0.003	0.088
$\sqrt{\text{var}(u_i)}$	σ	8.74	0.54				

For these data, subjects are highly heterogeneous ($\hat{\sigma} = 8.74$). Thus, strong associations exist among responses for the three situations. This is reflected by 1595 of the 1850 subjects making the same response on all three: that is, response patterns (0, 0, 0) and (1, 1, 1). This implies tremendous variability in between-subject odds ratios. From (13.8), for different subjects of a given gender, the middle 50% of odds ratios comparing situations 1 and 3 are estimated to vary between about $\exp(0.835 - 0.95 \times 8.74)$ and $\exp(0.835 + 0.95 \times 8.74)$.

For such contingency table data, finding cell fitted values requires integrating over the estimated random effects distribution to obtain estimated marginal probabilities of any particular sequence of

responses. For the ML parameter estimates, the probability of a particular sequence of responses (y_{i1}, \dots, y_{it}) for a given u_i is the appropriate product of conditional probabilities, $\prod_t P(Y_{it} = y_{it}|u_i)$, since the responses are assumed to be independent given u_i . Integrating this product probability with respect to u_i for the $N(0, \sigma^2)$ distribution estimates the marginal probability for a given cell (averaged over subjects). This requires numerical integration methods described in Section 13.6. Multiplying this marginal probability of a given sequence by the sample size for that multinomial gives a fitted value. For instance, of the females, 440 indicated support under all three circumstances (457 under none of the three), and the fitted value was 436.5 (459.3).

Overall chi-squared statistics comparing the 16 observed and fitted counts are $G^2 = 23.2$ and $X^2 = 27.8$ ($df = 9$). These are large, but the sample size is very large. Here, $df = 9$ since we are modeling 14 multinomial parameters ($8 - 1 = 7$ for each gender) using five GLMM parameters ($\alpha, \beta_2, \beta_3, \gamma, \sigma$). An extended model allows interaction between gender and situation. It has different $\{\beta_t\}$ for men and women. However, it does not fit better. The likelihood-ratio statistic comparing the models equals 1.0 ($df = 2$).

An alternative analysis of these data focuses on the marginal distributions, treating the dependence as a nuisance. A marginal model analog of (13.11) is

$$\text{logit}[P(Y_t = 1)] = \alpha + \beta_t + \gamma x.$$

For it, Table 13.4 also shows the ML estimates and the GEE estimates for the exchangeable working correlation structure. The marginal model fits well, with $G^2 = 1.10$; here, $df = 2$ since the model describes six marginal probabilities (three for each gender) using four parameters. These population-averaged $\{\beta_t\}$ are much smaller than the subject-specific $\{\beta_t\}$ from the GLMM. This reflects the very large GLMM heterogeneity ($\hat{\sigma} = 8.74$) and the corresponding strong correlations among the three responses. For instance, the GEE analysis estimates a common correlation of 0.82 between pairs of responses. Although the GLMM $\{\beta_t\}$ are about five to six times the marginal model $\{\beta_t\}$, so are the standard errors. The two approaches provide similar substantive interpretations and conclusions.

13.3.3 Example: Longitudinal Mental Depression Study Revisited

We now revisit [Table 12.1](#) from a longitudinal study to compare a new drug with a standard for treating subjects suffering mental depression. In Sections 12.1.1 and 12.2.2 we analyzed the data using marginal models. A response y_t on mental depression at time t equals 1 for normal and 0 for abnormal. For severity of initial diagnosis s (1 = severe, 0 = mild), drug treatment d (1 = new, 0 = standard), and t , we used the model

$$\text{logit}[P(Y_t = 1)] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 (d \times t)$$

to evaluate the marginal distributions.

Now let y_{it} denote the response at time t for subject i . The GLMM

$$\text{logit}[P(Y_{it} = 1|u_i)] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 (d \times t) + u_i$$

has subject-specific rather than population-averaged effects. [Table 13.5](#) shows the ML estimates. The time trend estimates of $\hat{\beta}_3 = 0.48$ for the standard drug and $\hat{\beta}_3 + \hat{\beta}_4 = 1.50$ for the new drug are nearly identical to the ML and GEE estimates (also shown in the table) for the corresponding marginal model. The reason is that the repeated observations do not exhibit much correlation, as the GEE analysis observed. Here, this is reflected by $\hat{\sigma} = 0.07$, showing little heterogeneity among subjects.

Table 13.5 Model Parameter Estimates for Marginal Model and Random Effects Logistic GLMM Fitted to [Table 12.1](#)

Parameter	Marginal ML		Marginal GEE		GLMM ML	
	Estimate	SE	Estimate	SE	Estimate	SE
Diagnosis	-1.29	0.14	-1.31	0.15	-1.32	0.15
Drug	-0.06	0.22	-0.06	0.23	-0.06	0.22
Time	0.48	0.12	0.48	0.12	0.48	0.12
Drug × Time	1.01	0.18	1.02	0.19	1.02	0.19

Based on the GLMM fit, integrating over the $N(0, 0.07^2)$ random effects distribution yields marginal fitted values of the possible response sequences. Comparing these to the sample counts in [Table 12.1](#) indicates a relatively good fit. The model describes the 28 multinomial cell probabilities (seven for the trivariate response at each of the four severity \times drug combinations) using six parameters. The fit statistics comparing the observed cell counts to their fitted values are $G^2 = 22.0$ and $X^2 = 20.8$ ($\text{df} = 28 - 6 = 22$).

The deviance increases by only 0.001 when we constrain $\sigma = 0$. From results to be discussed in Section 13.6.5, the P -value for comparing models is half what one gets by treating the deviance as chi-squared with $\text{df} = 1$. So, $P = 0.49$. This simpler model, which gives nearly identical effect estimates and SE values, is adequate. This is also suggested by AIC values (e.g., PROC NLMIXED in SAS reports 1173.9 for the GLMM and 1171.9 for the simpler model with $\sigma = 0$).

13.3.4 Example: Capture–Recapture Prediction of Population Size

Capture–recapture experiments use a series of samples to estimate the size of a population. Such experiments have traditionally been used to estimate animal abundance in some habitat. At each sampling occasion, animals are captured and marked in some manner. The animals captured for any given sample are freed and all animals are candidates for recapture in a later sample. With T sampling occasions, a 2^T contingency table displays the data, with scale (captured, not captured) at each occasion. The count $n_{22\dots 2}$ is missing for the cell corresponding to noncapture at each occasion. If we knew this cell count, adding it to the others would yield the population size. Models specified for this 2^T table use the $2^T - 1$ observed counts to fit the model. The fit refers to those $2^T - 1$ cells, but extrapolating it yields an estimated count in the unobserved cell. Adding that to the total of the observed counts yields an estimate of population size.

To illustrate, with $T = 2$ captures, we observe n_{11} animals at both occasions, n_{12} at the first but not the second, and n_{21} at the second but not the first. We do not know the number n_{22} not captured either time. If we assumed independence in the 2×2 table, the prediction \hat{n}_{22} would be the value giving an odds ratio of 1.0; but $(n_{11}\hat{n}_{22})/(n_{12}n_{21}) = 1$ implies that $\hat{n}_{22} = n_{12}n_{21}/n_{11}$. This yields a population size prediction of

$$\begin{aligned}\hat{N} &= n_{11} + n_{12} + n_{21} + n_{12}n_{21}/n_{11} \\ &= n_{1+}n_{+1}/n_{11} \quad \text{with} \quad \widehat{\text{var}}(\hat{N}) = \frac{n_{1+}n_{+1}n_{12}n_{21}}{n_{11}^3}\end{aligned}$$

(Sekar and Deming 1949). The assumption of independence is usually unrealistic, however. With additional sampling occasions, we can base our prediction on more complex models.

[Table 13.6](#), from Cormack (1989), refers to a study having $T = 6$ consecutive trapping days for a population of snowshoe hares. The study observed 68 hares. For instance, the table indicates that 3 hares were observed on the first day but on none of the other days. For simplicity, models for studies over a brief time period assume that no deaths, births, or immigration into the population occurred during the study period. This is called a *closed population*.

Table 13.6 Capture–Recapture Results for Snowshoe Hares

Capture 6	Capture 5	Capture 4	Capture 3, Capture 2, Capture 1								
			000	001	010	011	100	101	110	111	
0	0	0	—	3	6	0	5	1	0	0	
			(24.0) ^a	(2.3)	(5.4)	(0.9)	(3.2)	(0.5)	(1.2)	(0.3)	
0	0	1	3	2	3	0	0	1	0	0	
			(4.8)	(0.8)	(1.8)	(0.5)	(1.1)	(0.3)	(0.6)	(0.3)	
0	1	0	4	2	3	1	0	1	0	0	
			(3.9)	(0.6)	(1.5)	(0.4)	(0.9)	(0.2)	(0.5)	(0.2)	
0	1	1	1	0	0	0	0	0	0	0	
			(1.3)	(0.3)	(0.8)	(0.3)	(0.5)	(0.2)	(0.4)	(0.3)	
1	0	0	4	1	1	1	2	0	2	0	
			(6.8)	(1.1)	(2.6)	(0.6)	(1.5)	(0.4)	(0.9)	(0.4)	
1	0	1	4	0	3	0	1	0	2	0	
			(2.3)	(0.6)	(1.3)	(0.5)	(0.8)	(0.3)	(0.7)	(0.4)	
1	1	0	2	0	1	0	1	0	1	0	
			(1.9)	(0.5)	(1.1)	(0.4)	(0.7)	(0.3)	(0.6)	(0.4)	
1	1	1	1	1	1	0	0	0	1	2	
			(1.0)	(0.4)	(0.9)	(0.5)	(0.5)	(0.3)	(0.7)	(0.7)	

^aFitted values for logistic-normal model; 1 = capture, 0 = noncapture.

Source: Coull and Agresti (1999).

Most methods for capture–recapture treat the probability of capture at a given occasion as identical for each subject (e.g., animal). This is usually unrealistic. To allow heterogeneous capture probabilities, we use a logistic random effects model. For subject i , $i = 1, \dots, N$ with N unknown, let

$\mathbf{y}_i^T = (y_{i1}, \dots, y_{it})$, where $y_{it} = 1$ denotes capture in sample t and $y_{it} = 0$ denotes noncapture. Lacking explanatory variables, we could use the Rasch-type model

$$(13.12) \text{logit}[P(Y_{it} = 1|u_i)] = \alpha + \beta_t + u_i,$$

where $\{u_i\}$ are independent $N(0, \sigma^2)$, with a constraint such as $\beta_1 = 0$. The larger the value of β_t , the greater the capture probability at occasion t . The larger is σ , the more heterogeneous are the capture probabilities. When $\sigma = 0$ this logistic-normal model simplifies to mutual independence [i.e., loglinear model (9.6)] for the 2^T table.

As with other GLMMs, integrating the random effect from the probability mass function of $(\mathbf{y}_i|u_i)$ yields the likelihood function. We can consider this likelihood function and the resulting ML estimates of $\{\beta_t\}$ and σ for all possible counts in the unobserved cell. A profile likelihood function views the maximized likelihood as a function of the unobserved cell count. The ML prediction for that unobserved cell count is the value that maximizes this profile likelihood. Lacking specialized software, we can fit the GLMM repeatedly with various counts in the unobserved cell to determine by trial and error the count that maximizes the likelihood function. ML fitting to [Table 13.6](#) yields a prediction of 24.0 for the unobserved cell count. Since the study observed 68 hares, the population size estimate is $\hat{N} = 92$. For this fit, $\hat{\sigma} = 1.0$.

Methods for obtaining a confidence interval for N include using the profile likelihood function or a nonparametric bootstrap method. With the profile likelihood approach, the interval for the missing cell count consists of the possible counts for that cell such that the G^2 fit statistic increases by less than $\chi^2_1(\alpha)$ from its value at the ML estimate. Adding the number of subjects observed in the samples to the endpoints of this interval gives the corresponding interval for N . For the snowshoe hares, a 95% profile likelihood confidence interval for N is (75, 154). It is common for \hat{N} to be nearer the low end of the interval. See Coull and Agresti (1999) for details.

The greater the heterogeneity, as reflected by larger $\hat{\sigma}$, \hat{N} tends to be larger and the confidence interval tends to be wider. Large $\hat{\sigma}$ causes difficulties in estimation, since it results in a relatively flat likelihood surface. This implies imprecise estimates of N . In particular, the upper limit of the profile likelihood confidence interval for N is essentially infinite when the likelihood function gets sufficiently flat. Also, the ML estimator is then often unstable, with small changes in the data yielding large changes in \hat{N} . Difficulties can also arise when probabilities of capture are small. Evidence of this occurs when most subjects captured appear in only one sample. When this happens or when $\hat{\sigma}$ is large, confidence intervals for N are necessarily very wide.

Alternative models are discussed in Section 14.1.4. Models that ignore likely heterogeneity can give unrealistically narrow confidence intervals for N . Although traditionally used for animal populations, capture–recapture applications also include estimating population size for human populations. Darroch et al. (1993) considered census population estimation, and Chao et al. (2001) estimated the number of people infected during a hepatitis outbreak (Exercise 13.15). An interesting application is estimating the number of files on the World Wide Web relating to some subject by taking samples using several search engines (Fienberg et al. 1999).

13.3.5 Example: Heterogeneity Among Multicenter Clinical Trials

Many applications compare two groups on a categorical response for data stratified on a third variable. With binary outcomes, the data form several 2×2 contingency tables. The main focus relates to studying the association in the 2×2 tables and whether and how it varies among the strata.

The strata are sometimes themselves a sample, such as schools or medical clinics. A random effects approach is then natural. With a random sampling of strata, it enables inferences to extend to the population of strata. The fit of the random effects model provides a simple summary such as an estimated mean and standard deviation of log odds ratios for the population of strata. In each stratum it also provides a predicted log odds ratio that shrinks the sample value toward the mean. This is especially useful when the sample size in a stratum is small and the sample log odds ratio has large standard error. Even when the strata are not a random sample or not even a sample and a random effects approach is not as natural, the model is beneficial for these purposes.

We illustrate using [Table 13.7](#), previously analyzed in Section 6.4, showing the results of a clinical trial at eight centers. The purpose was to compare an active drug and a control, for curing a fungal infection. For a subject in center i using treatment t ($1 = \text{active drug}; 2 = \text{control}$), let $y_{it} = 1$ denote success. One possible model is the logistic-normal,

Table 13.7 Clinical Trial Relating Treatment to Response for Eight Centers

Center	Treatment	Response		Sample Odds Ratio	GLMM Fitted Odds Ratio
		Success	Failure		
1	Drug	11	25	1.19	2.02
	Control	10	27		
2	Drug	16	4	1.82	2.09
	Control	22	10		
3	Drug	14	5	4.80	2.19
	Control	7	12		
4	Drug	2	14	2.29	2.11
	Control	1	16		
5	Drug	6	11	∞	2.18
	Control	0	12		
6	Drug	1	10	∞	2.12
	Control	0	10		
7	Drug	1	4	2.00	2.11
	Control	1	8		
8	Drug	4	2	0.33	2.06
	Control	6	1		

Source: Beitler and Landis (1985).

$$\text{logit}[P(Y_{i1} = 1|u_i)] = \alpha + \beta/2 + u_i,$$

$$(13.13) \quad \text{logit}[P(Y_{i2} = 1|u_i)] = \alpha - \beta/2 + u_i,$$

where $\{u_i\}$ are independent $N(0, \sigma^2)$ variates. This model assumes that the log odds ratio β between treatment and response is constant over centers. The parameter σ summarizes center heterogeneity in the success probabilities.

A logistic-normal model permitting treatment-by-center interaction is

$$\text{logit}[P(Y_{i1} = 1|u_i, b_i)] = \alpha + (\beta + b_i)/2 + u_i,$$

$$(13.14) \quad \text{logit}[P(Y_{i2} = 1|u_i, b_i)] = \alpha - (\beta + b_i)/2 + u_i,$$

where $\{(u_i, b_i)\}$ are independent bivariate normal, with variances σ_u^2 and σ_b^2 and correlation ρ . The log odds ratio equals $\beta + b_i$ in center i . These vary among centers according to a $N(\beta, \sigma_b^2)$ distribution. That is, β is the expected center-specific log odds ratio between treatment and response, and σ_b describes variability in those log odds ratios. The model parameters are $(\alpha, \beta, \sigma_u, \sigma_b, \rho)$.

In [Table 13.7](#) the sample success rates vary markedly among centers both for the control and drug treatments, but in all except the last center that rate is higher for the drug treatment. In using models with random center and random treatment effects, it is preferable to have many more than eight centers. It is difficult to get reliable variance component estimates with so few centers, and the asymptotics for estimating the parameters for this type of model apply as the number of centers increases. Keeping this in mind, we use these data to illustrate the models. Simplifying the model a bit by taking $\rho = 0$ (justified by an ML estimate that is not significantly nonzero), the treatment estimates are $\hat{\beta} = 0.739$ ($SE = 0.300$) for the model [\(13.13\)](#) of no interaction and $\hat{\beta} = 0.746$ ($SE = 0.325$) for the model [\(13.14\)](#) permitting interaction. Considerable evidence of a drug effect occurs.

The evidence about association is weaker for the model permitting interaction. The Wald statistics are $(0.739/0.300)^2 = 6.0$ for the no-interaction model and $(0.746/0.325)^2 = 5.3$ for the interaction model ($df = 1$). The corresponding likelihood-ratio statistics are 6.3 and 4.6. The extra variance component in the interaction model pertains to variability in the log odds ratios. As its estimate $\hat{\sigma}_b$ increases, so does the SE of the estimated treatment effect $\hat{\beta}$ tend to increase. In this example, $\hat{\sigma}_b = 0.15$ is relatively small and the standard errors of $\hat{\beta}$ are not very different in the two models. When $\hat{\sigma}_b = 0$, the standard errors and the model fits are the same.

To show the effect of larger $\hat{\sigma}_b$ on the SE of the mean treatment effect estimate $\hat{\beta}$, we alter [Table 13.7](#) slightly. We change three failures to successes for drug in center 3 and three successes to failures for drug in center 8. With these changes, the estimated variability of the treatment effects increases from $\hat{\sigma}_b = 0.15$ to $\hat{\sigma}_b = 1.37$. The ML estimates of the mean treatment effects are then $\hat{\beta} = 0.722$ ($SE = 0.299$) for the no-interaction model [\(13.13\)](#) and $\hat{\beta} = 0.767$ ($SE = 0.623$) for the interaction model. The Wald statistics are 5.84 and 1.52. The evidence of a treatment effect is then dramatically weaker for the interaction model. Not surprisingly, when the treatment effect varies substantially among centers, it is more difficult to estimate the mean of that effect.

For the actual data in [Table 13.7](#), because $\hat{\sigma}_b = 0.15$ for model [\(13.14\)](#) is relatively small, the model shrinks the sample odds ratios considerably. [Table 13.7](#) shows the sample values and the model predicted values. These are based on predicting the random effects and substituting them and the ML estimates of fixed effects into the model formula to estimate the two response probabilities for each treatment in each center. The sample odds ratios vary from 0.33 to ∞ ; their GLMM counterparts vary only between 2.02 and 2.19. The smoothed estimates are much less variable and do not have the same ordering as the sample values. For instance, the smoothed estimate of 2.19 for center 3 is greater than the estimate of 2.12 for center 6, even though the sample value is infinite for the latter. This reflects the greater shrinkage that occurs when sample sizes are smaller.

13.3.6 Meta-analysis Using a Random Effects Approach

In Section 6.4.6 we discussed ways of summarizing multiple 2×2 tables, such as arise in meta-analyses to compare two treatments on a binary response. The analyses just shown for multicenter clinical trials also apply naturally to meta-analyses. The models in Section 6.4.6 did not allow for heterogeneity in the effects or allow for treating the studies as a random sample of potential studies. With random effects models, we have a natural way to do both of these.

The model (13.14) provides a summary of variability in log odds ratios among studies. For simpler interpretation, we might want to describe variability among relative risk or difference of proportion values. This can be done with analogous models using log or identity links. However, there are structural problems, as the linear predictor can take any possible real number value when it contains a normal random effect. Dersimonian and Laird (1986) proposed a random effects approach with the difference of proportions, treating the difference of proportions as coming from a normal distribution. Their method uses weighted least-squares estimators with weights based on sample estimates of variances of proportions. This approach, compared with ML, can behave poorly for small samples with true proportions near the boundary, and it can be quite biased when the weights are correlated with the study effect sizes of interest. Warn et al. (2002) used a Bayesian approach, discussed in Section 13.7.2, that imposes constraints that reflect the bounds for probabilities.

A challenging situation for meta-analyses is when the outcome of interest has very low probability. Some tables may have empty cells for one or both treatments. As discussed in Section 6.4.6, such cases do not affect statistical significance but do provide evidence about the magnitude of the difference of proportions and its variability. See Emerson et al. (1993) and references in the Laird et al. comments about Shuster (2010).

13.3.7 Alternative Formulations of Random Effects Models

There are other ways to express random effects models. For instance, an equivalent expression for interaction model (13.14) is

$$\text{logit}[P(Y_{it} = 1|u_i, b_{it})] = \alpha + \beta x_t + b_{it} + u_i,$$

where x_t is a treatment indicator variable ($x_1 = 1, x_2 = 0$). Here, $b_{i1} - b_{i2}$ corresponds to b_i in parameterization (13.14), and $2\sigma_b^2$ here corresponds to σ_b^2 in (13.14).

Formulating a random effects model requires care about implications of the model expression and the random effects correlation structure. Suppose we expressed this interaction model as

$$(13.15) \quad \text{logit}[P(Y_{it} = 1|u_i, b_i)] = \alpha + (\beta + b_i)x_t + u_i,$$

with $\{b_i\}$ from $N(0, \sigma_b^2)$ independent of $\{u_i\}$ from $N(0, \sigma_u^2)$. This is inappropriate, because the model then imposes greater variability for the logit with the first treatment than the second, since $x_2 = 0$ and $\{u_i\}$ and $\{b_i\}$ are uncorrelated. Also, the model should not depend on the definition of the indicator variable x_t . Note, however, that if $z_t = x_t + c$ for some constant c , then model (13.15) is equivalently

$$\text{logit}[P(Y_{it} = 1|u_i, b_i)] = \alpha + (\beta + b_i)(z_t - c) + u_i = \alpha' + (\beta + b_i)z_t + v_i,$$

where $\alpha' = \alpha - c\beta$ and $v_i = u_i - cb_i$. Thus, (v_i, b_i) are correlated even if (u_i, b_i) are not. In fact, expression (13.15) is sensible only with correlated random effects. It is then equivalent to (13.14) with correlated random effects.

Rabe-Hesketh and Skrondal (2001) showed that careful attention must be paid to parameter identification in models with multivariate random effects. Their *factor model* contains many multivariate random effects models as special cases.

13.3.8 Example: Matched Pairs with a Bivariate Binary Response

A sample of schoolboys were interviewed twice, several months apart, and asked about their self-perceived membership in the “leading crowd” and about whether they sometimes needed to go against their principles to belong to that group. Thus, there are two binary response variables, which we refer to as membership and attitude, measured at two interview times for each subject. [Table 13.8](#) labels the categories for attitude as (positive, negative), where “positive” refers to disagreeing with the statement that one must go against his principles.

Table 13.8 Membership and Attitude Toward the “Leading Crowd”

(M, A) for First Interview	(M, A) for Second Interview ^a			
	(Yes, Positive)	(Yes, Negative)	(No, Positive)	(No, Negative)
Yes, positive	458	140	110	49
Yes, negative	171	182	56	87
No, positive	184	75	531	281
No, negative	85	97	338	554

^a M, membership; A, attitude.

Source: J. S. Coleman, *Introduction to Mathematical Sociology*. London: Free Press of Glencoe, 1964, p. 170.

For subject i , let y_{itv} be the response at interview time t on variable v , where $v = M$ for membership and $v = A$ for attitude. The logistic model

$$(13.16) \text{ logit}[P(Y_{itv} = 1|u_{iv})] = \alpha + \beta_{tv} + u_{iv}$$

is a multivariate form of the Rasch-type model (13.4). It has additive item and subject effects for each variable v . Here, (u_{iM}, u_{iA}) is a bivariate random effect that describes subject heterogeneity for (membership, attitude). We assume that the $\{(u_{iM}, u_{iA})\}$ are independent from a bivariate normal distribution, $N(\mathbf{0}, \Sigma)$, with possibly different variances and nonzero correlation.

The ML fit yields $\hat{\beta}_{2M} - \hat{\beta}_{1M} = 0.379$ ($SE = 0.075$) and $\hat{\beta}_{2A} - \hat{\beta}_{1A} = 0.176$ ($SE = 0.058$). For both variables, the probability of the first outcome category is higher at the second interview. For instance, for a given subject the odds of self-perceived membership in the leading crowd at interview 2 are estimated to be $\exp(0.379) = 1.46$ times the odds at interview 1.

The estimated correlation between the random effects is 0.32. Their estimated standard deviations are $\hat{\sigma}_1 = 3.08$ for $\{u_{iM}\}$ and $\hat{\sigma}_2 = 1.49$ for $\{u_{iA}\}$. Since these are quite different, the relative sizes of membership and attitude effects differ for marginal and random effects models (recall the caveat in Section 13.2.4). The marginal effect is attenuated more for membership. For this random effects model, the ratio of estimated odds ratios is $\exp(0.379)/\exp(0.176) = 1.46/1.19 = 1.22$. For the marginal model, the estimated odds ratios use the marginal distributions of each variable at each time [e.g., this is $(1392/2006)/(1253/2145) = 1.188$ for membership], and the ratio of estimated odds ratios is $1.188/1.133 = 1.05$.

Integrating over the estimated random effects distribution yields fitted values for the 16 possible sequences of responses in [Table 13.8](#). The deviance of $G^2 = 5.5$ ($df = 8$) compares the 16 observed counts to their fitted values. The model, which describes 15 multinomial probabilities with seven parameters, fits well. The model constraining the random effects to be uncorrelated fits poorly ($G^2 = 97.5$, $df = 9$). The model constraining the random effects to be perfectly correlated is equivalent to having a single random effect u_i for each subject. The model is then a Rasch-type model with four items that are the combinations of interviews and variables. That model fits very poorly ($G^2 = 655.5$, $df = 10$).

13.3.9 Time Series Models Using Autocorrelated Random Effects

Section 12.4 noted that categorical time series data can be modeled with transitional models in which previous response values as well as ordinary explanatory variables serve as predictors in the model for Y_t . When a main purpose is to describe the effect of an explanatory variable x_t on $E(Y_t)$, a disadvantage of such models is that the interpretation of the β coefficient of x_t depends on how many previous response values are in the model. For a first-order Markov logistic model, for instance, β refers to the impact on $\text{logit}[P(Y_t = 1)]$ of a 1-unit increase in x_t , but at a fixed value of y_{t-1} .

In an alternative approach, Klingenberg (2008) proposed GLMMs in which the serial dependence is accounted for by random effects having an autoregressive structure. For binary data, generalizing the logistic-normal model (13.6), he assumed that

$$(13.17) \quad \text{logit}[P(Y_{it} = 1|u_t)] = \mathbf{x}_{it}^T \boldsymbol{\beta} + u_t,$$

where

$$u_t = \rho u_{t-1} + \epsilon_t, \quad t = 2, 3, \dots, T,$$

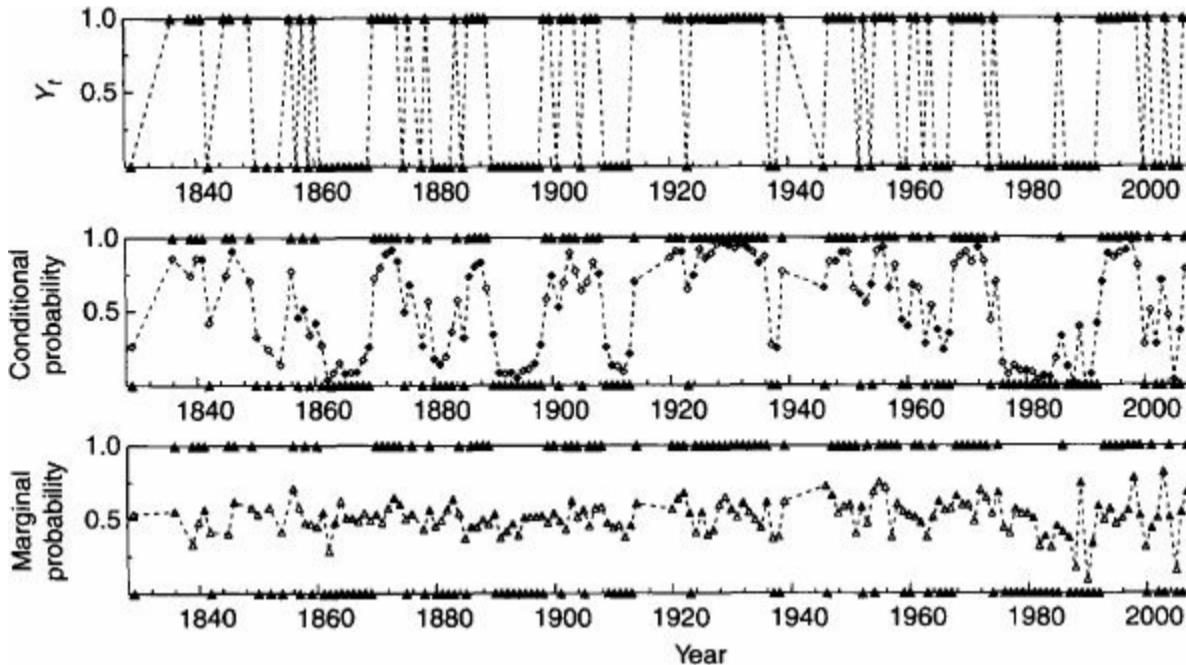
and $\{\epsilon_t\}$ are uncorrelated normal variates. He parameterized by setting $u_1 \sim N(0, \sigma^2)$ and taking $\epsilon_t \sim N(0, \sigma^2[1 - \rho^2])$, so that $\text{var}(u_t) = \sigma^2$ for all t . More generally, there can be a varying time lag d_t between y_{t-1} and y_t in which case ρ is replaced by ρ^{d_t} .

In model (13.17), the effect of an explanatory variable is conditional on the random effect but not on previous response values. ML model fitting is complex, because the random effect to be integrated out to obtain the likelihood function is T -dimensional. Klingenberg (2008) presented a Monte Carlo EM algorithm for doing this.

13.3.10 Example: Oxford and Cambridge Annual Boat Race

Klingenberg illustrated the time series model using the outcomes of 152 races between the rowing teams from Cambridge and Oxford, for the famous race held nearly every year since 1829. Let $y_t = 1$ when Cambridge wins and $y_t = 0$ when Oxford wins. [Figure 13.2](#) shows the data, which are available at www.theboatrace.org.

[Figure 13.2](#) Results of annual rowing race (1 = Cambridge wins, 0 = Oxford wins) in first panel, with estimated conditional (second panel) and marginal (third panel) probabilities of a Cambridge win. Source: Klingenberg (2008). Used with permission.



A 2×2 table cross-classifying y_t by y_{t-1} for the 151 adjacent pairs of races has counts 48 for (1,1), 43 for (0,0), and 25 for both (1,0) and (0,1), giving an odds ratio of 3.30 between successive responses. Klingenberg focused on whether the weight differential between the crews has an effect. Let x_t be the average weight difference, in pounds per crewman, between the Cambridge team and the Oxford team in year t . The model

$$\text{logit}[P(Y_t = 1|u_t)] = \alpha + \beta x_t + u_t$$

with autoregressive normal random effect has $\hat{\alpha} = 0.25$ ($SE = 0.44$), $\hat{\beta} = 0.14$ ($SE = 0.06$), $\hat{\sigma} = 2.03$ ($SE = 0.81$), and $\hat{\rho} = 0.69$ ($SE = 0.12$). Weight seems to have a positive effect, but the estimated size of that effect is rather imprecise. The substantial $\hat{\rho}$ value reflects the strong association between successive outcomes.

For his model fit, [Figure 13.2](#) also shows the estimated conditional and marginal probabilities of a Cambridge win. The influence of the predicted random effect is to move many of the estimated conditional probabilities well away from 0.50, compared with the estimated marginal probabilities. Klingenberg checked the fit of the model in various ways. One way used the model fit to estimate three-dimensional transition probabilities for sequences of three races in a row, and compared these to the observed proportions for those sequences. The fit seemed to be quite good.

13.4 RANDOM EFFECTS MODELS FOR MULTINOMIAL DATA

Random effects models for binary responses extend to multicategory responses. For the multicategory models of Chapter 8, adding random effects extends this multivariate GLM to a multivariate GLMM (Hartzel et al. 2001b). This class includes models for nominal and ordinal responses.

13.4.1 Cumulative Logit Model with Random Intercept

Modeling is simpler with ordinal than nominal responses, since often the same random effect and the same fixed effect can apply to each logit. With cumulative logits, this is the *proportional odds* structure (Section 8.2.2). Denote the possible outcomes for y_{it} , observation t in cluster i , by 1, 2, ..., I . A GLMM for the cumulative logits has the form

$$(13.18) \text{logit}[P(Y_{it} \leq j | \boldsymbol{u}_i)] = \alpha_j + \mathbf{x}_{it}^T \boldsymbol{\beta} + \mathbf{z}_{it}^T \boldsymbol{u}_i, \quad j = 1, \dots, I - 1.$$

Hedeker and Gibbons (1994) and Tutz and Hennevogl (1996) discussed model fitting, primarily by treating \boldsymbol{u}_i as multivariate normal.

For cumulative logit and probit random intercept models, the same relationship exists between their effects and those in marginal models as presented in Section 13.2.3 for binary-response models. Marginal effects tend to be smaller, increasingly so as σ increases.

13.4.2 Example: Insomnia Study Revisited

[Table 12.3](#) showed results of a clinical trial at two occasions comparing a drug with placebo in treating insomnia patients. In Sections 12.1.3 and 12.2.3 we analyzed the data with marginal models. For $y_t = \text{time to fall asleep at occasion } t$, the marginal model

$$\text{logit}[P(Y_t \leq j)] = \alpha_j + \beta_1 t + \beta_2 x + \beta_3(t \times x)$$

permits interaction between $t = \text{occasion}$ (0 = initial, 1 = follow-up) and $x = \text{treatment}$ (1 = active, 0 = placebo). [Table 13.9](#) shows the ML and GEE estimates.

Table 13.9 Fits of Cumulative Logit Models to Insomnia Data in [Table 12.3](#)^a

Effect	Marginal ML	Marginal GEE	Random Effects (GLMM) ML
Treatment	0.046 (0.236)	0.034 (0.238)	0.058 (0.366)
Occasion	1.074 (0.162)	1.038 (0.168)	1.602 (0.283)
Treatment \times occasion	0.662 (0.244)	0.708 (0.244)	1.081 (0.380)

^aValues in parentheses are standard errors.

Now, let y_{it} denote the response for subject i at occasion t . [Table 13.9](#) also shows results of fitting the random intercept model

$$\text{logit}[P(Y_{it} \leq j|u_i)] = \alpha_j + \beta_1 t + \beta_2 x + \beta_3(t \times x) + u_i.$$

Results are substantively similar to the marginal model, but estimates and standard errors are about 50% larger. This reflects the relatively large heterogeneity ($\hat{\sigma} = 1.90$) and the resultant strong association between the responses at the two occasions.

13.4.3 Example: Combining Measures on Ordinal Items

In many surveys, subjects respond to a set of items that measure various aspects of some characteristic, each using the same ordinal scale. For example, some quality of life instruments have separate questions pertaining to the frequency with which a person participates in various activities, with each activity measured with a scale such as (never, rarely, occasionally, often). In such cases with multiple response items of a similar nature, it can be useful to combine the response outcomes into a single score for each subject that provides a summary measure of that characteristic.

One approach, commonly used, is to assign scores such as (1,2,3,4) to the outcome categories and for each subject find the mean score on the set of items. This has the advantage of simplicity of calculation and of interpretation, but it is often not obvious how to assign the scores, such as with the quality of life scale just mentioned. An alternative way of forming a summary measure mimics item-response modeling for binary data (Section 13.1.4).

To illustrate, we use some General Social Survey data. The GSS often asks subjects their opinion about government spending in various areas, such as health, education, the environment, culture and the arts, defense, law enforcement, unemployment benefits, and retirement benefits. The outcome scale is (much more, more, the same, less, much less). We use only three items here and combine categories 1 and 2 and combine categories 4 and 5 so we can easily show the data, which are in [Table 13.10](#).

Table 13.10 Opinions About Government Spending on the Environment, Education, and Health (1 = more, 2 = same, 3 = less)

Education	Environment = 1 Health			Environment = 2 Health			Environment = 3 Health		
	1	2	3	1	2	3	1	2	3
1	651	45	15	304	59	10	92	24	17
2	57	10	3	50	35	12	15	14	6
3	7	1	5	7	10	4	6	3	16

Source: 2006 General Social Survey.

For subject i with item t ($1 = \text{education}$, $2 = \text{health}$, $3 = \text{environment}$), we use the random intercept model

$$\text{logit}[P(Y_{it} \leq j|u_i)] = \alpha_j + \beta_1 I(t = 1) + \beta_2 I(t = 2) + u_i,$$

were $I(t = 1)$ and $I(t = 2)$ are indicators for the first two items. Here, u_i reflects the propensity of subject i to be favorable to more government spending, relatively greater values increasing the chance of response at the low end of the scale corresponding to higher spending. The ML fit of this model has $\hat{\beta}_1 = 1.888$ ($SE = 0.102$), $\hat{\beta}_2 = 1.540$ ($SE = 0.086$), and $\hat{\sigma} = 1.56$ ($SE = 0.08$), reflecting a tendency for subjects to prefer less spending on the environment than the other two items. The latent variable structure that generates this form of model (Section 8.2.3) suggests that differences between pairs of $\{u_i\}$ are location shifts between subjects in the distribution of a latent variable for government spending, with other shifts (described by $\{\beta_t\}$) according to the item considered.

To construct a summary measure for the subjects that reflects their propensities to be favorable to more government spending, we predict $\{u_i\}$ using their posterior means based on the ML model fit. To illustrate, [Table 13.11](#) shows these for some of the possible response sequences. Note that compared with the scores obtained by assigning fixed scores to the categories and finding the mean outcome for each subject:

Table 13.11 Cumulative Logit Predicted Random Effect Values for Opinions About Government Spending Data in [Table 13.10](#)

Envir.	Educ.	Health	\hat{u}_i Predictions	Envir.	Educ.	Health	\hat{u}_i Predictions
1	1	1	0.88	2	1	1	-0.37
1	1	2	-0.31	2	2	2	-1.78
1	2	1	-0.32	1	2	3	-1.62

- For a given set of response outcomes, a subject's predicted score is the same for any permutation of the outcomes for the fixed scores approach, but not for the cumulative logit model or models with other links. For example, note in [Table 13.11](#) the results for response sequences (1,1,2), (1,2,1), and (2,1,1).
- The model-based predicted scores are not a linear function of the mean of assigned fixed scores. The spacing is governed by the shape of the distribution for the underlying latent variable model, for example, logistic for cumulative logit link.

13.4.4 Example: Cluster Sampling

With surveys that use cluster sampling, standard methods based on simple random sampling (e.g., for a single multinomial sample) require adjustment. Ordinary standard errors are too small. When the sampling scheme randomly samples clusters, we can account for the clustering using cluster random effects. We illustrate using data from Brier (1980), who reported 96 observations taken from 20 neighborhoods (the clusters) on Y = satisfaction with home and x = satisfaction with neighborhood as a whole. The data are shown at the text website. Each variable was measured with the ordinal scale (unsatisfied, satisfied, very satisfied). Brier's analysis adjusted for clustering by reducing the Pearson statistic for testing independence in the 3×3 contingency table relating X and Y from 17.9 to 15.7 ($df = 4$).

Consider the model for y_{it} , observation t in cluster i ,

$$(13.19) \text{logit}[P(Y_{it} \leq j|u_i)] = \alpha_j + x_{it}\beta + u_i,$$

with scores (1, 2, 3) for the satisfaction levels of x_{it} . With a $N(0, \sigma^2)$ distribution assumed for u_i , the ML effect estimate is $\hat{\beta} = -1.201$ ($SE = 0.407$), with $\hat{\sigma} = 0.92$. By contrast, treating the 96 observations as a random sample corresponds to fitting this model with $\sigma = 0$. It has $\hat{\beta} = -1.226$ ($SE = 0.370$). A slight reduction in significance results from adjusting for clustering.

Rao and Thomas (1988) surveyed ways of adjusting standard inferences to take into account complex sampling methods in the analysis and modeling of categorical data. The usual chi-squared test statistics no longer have chi-squared null distributions, but rather, weighted sums of chi-squared. See also references in Note 3.4.

13.4.5 Baseline-Category Logit Models with Random Effects

For nominal response variables, we can formulate a binary GLMM that pairs each category with a baseline and fit these models simultaneously while allowing separate effects. This requires using a vector of cluster-specific random effects \mathbf{u}_{ij} , one for each logit. The general form of the baseline-category logit model with random effects is

$$\log \frac{P(Y_{it} = j)}{P(Y_{it} = I)} = \alpha_j + \mathbf{x}_{it}^T \boldsymbol{\beta}_j + \mathbf{z}_{it}^T \mathbf{u}_{ij}, \quad j = 1, \dots, I - 1.$$

The fixed effects $\boldsymbol{\beta}_j$ and the random effects \mathbf{u}_{ij} depend on j , since the baseline category is arbitrary.

Cluster i has a vector $\mathbf{u}_i^T = (\mathbf{u}_{i,I-1}^T)$ of random effects, treated as independent multivariate normal variates. We recommend an unspecified covariance matrix Σ , for \mathbf{u}_i , to allow different variances for random effects that apply to different logits. With a common variance, that variance would not be the same as that for the implied random effect for a logit for an arbitrary pair of categories, $\log[P(Y_{it} = j)/P(Y_{it} = k)]$. With unspecified covariance the model is structurally the same regardless of the choice of baseline category.

13.4.6 Example: Effectiveness of Housing Program

Hedeker (2008) discussed a California study² designed to investigate the effectiveness of a housing certificate regarding whether individuals diagnosed with mental illness who were homeless or at high risk of becoming homeless were able to choose and stay in independent housing in their community. The housing certificates required clients to pay 30% of their income toward rent. Eligible subjects were randomly assigned to two groups, one of which received the certificates and the other of which was a control group. Initially and after 6, 12, and 24 months the subjects' housing status was classified as (independent housing, community housing, streets/shelters).

Let c indicate whether a subject was in the certificate group ($1 = \text{yes}$, $0 = \text{no}$), and let t_1 , t_2 , and t_3 be indicators for contrasting each time with the initial baseline. For subject i at time t , Hedeker proposed the model

$$\log \frac{P(Y_{it} = j)}{P(Y_{it} = 3)} = \alpha_j + \beta_{1j}t_1 + \beta_{2j}t_2 + \beta_{3j}t_3 + \beta_{4j}c + \beta_{5j}(c \times t_1) \\ + \beta_{6j}(c \times t_2) + \beta_{7j}(c \times t_3) + u_{ij},$$

for $j = 1, 2$. [Table 13.12](#) shows ML estimates and SE values for the model, based on treating missing observations as missing at random.

Table 13.12 ML Estimates for Baseline-Category Logit Random Effects Model for Modeling Effectiveness of Program for Independent Housing

Effect	Independent vs. Street		Community vs. Street	
	Estimate	SE	Estimate	SE
Intercept	-2.67	0.37	-0.45	0.19
t_1 (6 months)	2.68	0.42	1.94	0.31
t_2 (12 months)	4.09	0.56	2.82	0.47
t_3 (24 months)	4.10	0.47	2.26	0.38
c (certificate indicator)	0.78	0.49	0.52	0.27
$c \times t_1$	2.00	0.61	-0.14	0.49
$c \times t_2$	0.55	0.69	-1.92	0.61
$c \times t_3$	0.30	0.62	-0.95	0.54
σ (random effects)	2.33	0.20	0.87	0.14

Source: With kind permission from Springer, from Tables 6.4 and 6.5 of Hedeker (2008).

From the first logit, the log odds ratio comparing the two certificate groups in terms of independent housing vs. street/shelters is 0.78 initially, $0.78 + 2.00 = 2.78$ after 6 months, $0.78 + 0.55 = 1.33$ after 12 months, and $0.78 + 0.30 = 1.08$ after 24 months. The increase in independent housing from the initial measurement is quite pronounced for the certificate group (relative to the control group) after 6 months, but is not significant at 12 or 24 months. The relatively large $\hat{\sigma} = 2.33$ for the random effects for this logit reflects strong positive associations among the repeated responses, conditional on response in one of these two categories.

We leave further interpretation of effects to Exercise 13.19. A more complex model permitting a different σ for each group for each logit did not fit significantly better.

13.5 MULTILEVEL MODELING

In some research studies, the data structure is hierarchical, with sampled units nested in clusters that are themselves nested in other clusters. Patients are nested in hospitals, which are themselves nested in communities. Individuals are nested in families. In a longitudinal study, repeated measurements are nested within a cluster that is a person observed over time. Random effects can enter models for such data at different levels of the hierarchy.

Many early uses of hierarchical models were in educational applications (e.g., Aitkin et al. 1981). A statewide study of factors that affect student performance might measure each student's scores on a battery of exams but use a model that takes into account the student, the school or school district, and the county. Just as two observations on the same student will tend to be more alike than observations on different students, so will two students in the same school tend to be more alike than two students from different schools. Student, school, and county terms can be treated as random effects, with different ones referring to different levels of the model. A model might have students at level 1, schools at level 2, and counties at level 3. GLMMs for data having a hierarchical structure of this sort are called *multilevel models*.

When the data have a hierarchical structure, the model should reflect that fact rather than ignore it. The researcher can then pay attention to explanatory variables that are relevant at each level, and decompose the total error variability into portions corresponding to each level. Using multilevel models also helps to correct for biases that would occur if we ignored the clustering and the consequent within-cluster correlations.

13.5.1 Hierarchical Random Terms: Partitioning Variability

We illustrate with a two-level model. Let $y_{i(j)t}$ denote the response for student i , who attends school j , on test t in a battery of tests, with 1 = pass and 0 = fail. A multilevel model with random effects $\{v_{i(j)}\}$ for students and $\{u_j\}$ for schools and fixed effects for explanatory variables has the form

$$(13.20) \quad \text{logit}[P(Y_{i(j)t} = 1)] = \mathbf{x}_{i(j)t}^T \boldsymbol{\beta} + u_j + v_{i(j)}.$$

Here, the explanatory variables x might include a factor that identifies the test in the battery (with categories such as math, verbal, ...). The random effects u_j and $v_{i(j)}$ are assumed to be independent with distributions $N(0, \sigma_u^2)$ and $N(0, \sigma_v^2)$ having unknown variances.

The level 1 random effects $\{v_{i(j)}\}$ account for variability among students in ability as well as in characteristics that are not measured in x , such as perhaps their achievement motivation. A relatively large σ_v value induces a strong correlation among the test results for students. The level 2 random effects $\{u_j\}$ account for variability among schools due to possibly unmeasured variables such as quality of the teachers. Model (13.20) is a random intercept model. More general models also can have random slopes for effects of explanatory variables.

As in Section 7.1.1, a latent variable model implies this model. Let $y_{i(j)t}^*$ denote the latent observation for student i in school j on test t , such that we observe $y_{i(j)t} = 1$ if $y_{i(j)t}^*$ falls above some threshold, such as 0. The latent model is

$$y_{i(j)t}^* = \mathbf{x}_{i(j)t}^T \boldsymbol{\beta} + u_j + v_{i(j)} + \epsilon_{i(j)t}.$$

The assumption that $\{\epsilon_{i(j)t}\}$ come from a standard logistic distribution, for which the inverse cdf is the logit link function, implies the logistic random effects model. For it, conditional on u_j and $v_{i(j)}$, the observed response satisfies the logistic model (13.20). The assumption that $\epsilon_{i(j)t}$ come from a standard normal distribution implies a corresponding probit random effects model.

Although the random effects enter at two levels, this representation shows that such a model actually has three levels: A particular observation is affected (beyond the influence of the explanatory variables) by random variability among schools, among students within the school, and among tests taken by a student.

With independent error terms, the total unexplained variability in this latent variable model is $\text{var}(u_j) + \text{var}(v_{i(j)}) + \text{var}(\epsilon_{i(j)t})$, where $\text{var}(\epsilon_{i(j)t}) = \pi^2/3 = 3.29$ for the logistic model and 1.0 for the probit model. Strong correlations between scores on different tests for the students correspond to a relatively large value of $\text{var}(v_{i(j)})$, and a relatively large proportion of the total variability that is due to this variability among students.

13.5.2 Example: Children's Care for an Unmarried Mother

What factors help to predict whether an adult child provides care for an unmarried elderly mother? This was recently investigated by sociologist J. Henretta and two colleagues using a longitudinal study with a cohort of subjects from the Health and Retirement Study. Their study used 16,719 observations on 5607 mother–child pairs in 1925 families. The outcome measure was whether a child provided care for the mother by providing financial help or assistance with daily living tasks. This was observed by interviews of the mothers in 1998, 2000, 2002, and 2004. The study was restricted to families in which the mother was unmarried at her first interview, since spouses rather than children are the primary helpers of married elders. About half the mothers died during the study period, and proxy respondents were used for one interview after this happened. Overall, care was reported to be provided in 18% of the observations.

Let $y_{i(j)t}$ denote the response for child i in family j about whether he/she provides care for their elderly mother when observed at time t (1 = yes, 0 = no). The explanatory variables were ethnicity, the year of the observation, characteristics of the mother (health, age, assets, whether the mother was experiencing her final illness before death), characteristics of the child (sex, whether married, whether a stepchild, whether has children, whether attended college, whether the mother raised a child of his/hers for at least a year, whether the child received from the mother at least \$5000 in financial help in the 10 years preceding 1993), and characteristics of the family (family size, % of children who are male, % of children who are married, % of children who are a stepchild, % of children who have their own children, % of children who attended college, whether the mother's family received financial help from relatives before she was of age 16). All these variables were categorical in measurement.

Positive correlation occurs among repeated observations over time for a child and also among observations from different children within the same family. Thus, the researchers posed multilevel models of form (13.20) with a random effect for each child at level 1 and a random effect for each family at level 2. Table 13.13 shows ML estimates and SE values for their full model. All multicategory explanatory variables were treated as nominal, using a set of indicator variables. For example, the estimates shown for assets used the sixth of the seven categories (namely, \$100,000–249,000) as the baseline, which was the most common category. The estimates indicate that a higher probability of help is associated with the mother having poorer health and more advanced age and in her final illness, with the child being female and not a stepchild and without children, with the family being smaller and with relatively more children who are male and who themselves have children, and when the mother reported that her family received help when she was growing up.

To illustrate interpretation of explanatory effects, consider final illness. For a given child and fixed values of other explanatory variables, the estimated odds of providing help during the mother's final illness were $\exp(1.411) = 4.1$ times the estimated odds when it was not that time. Henretta and colleagues also provided estimated probabilities of providing help. At the most common categories for each of the other explanatory variables and at random effect values of 0, the estimated probabilities were 0.17 when it was the time of the final illness and 0.05 when it was not. Such probability estimates vary considerably according to values of explanatory variables. For example, this pair (0.17, 0.05) of estimated probabilities for care under (final illness, not final illness) changed for cases of only biological children to (0.46, 0.17) for females and (0.34, 0.11) for males, and for biological children with two female sibs to (0.13, 0.03) for females and (0.05, 0.01) for males. The pairs show that the effect of final illness is very strong. These last four pairs also provide a description of the sex effect, with daughters being more likely to provide care, and show that care is much more likely by an only child than by a child with two female sibs.

For inference, from Table 13.13 it is possible only to construct Wald tests and confidence intervals comparing each category of a factor to its baseline. However, using log-likelihood values L reported by software for the model and for the model with the factor removed, it's possible to do the usual

likelihood-ratio tests. For example, Henretta and colleagues reported $L = -6133.2$ for the model shown in [Table 13.13](#) and $L = -6255.3$ for the simpler model that removes all the family characteristics. Double that difference, which equals 244.2, is a chi-squared statistic with $df = 12$ for testing the hypothesis that none of the family characteristics have an effect.

Table 13.13 ML Estimates and *SE* Values for Multilevel Model for Whether an Adult Child Cares for Her Unmarried Elderly Mother

Effect	Estimate	SE	Effect	Estimate	SE
Intercept	-2.027	0.317			
Ethnicity (vs. White)			<i>Child characteristics</i>		
Black	0.162	0.157	Sex (Male = 1)	-1.435	0.118
Hispanic	-0.165	0.207	Married (Yes = 1)	-0.179	0.119
Other	0.459	0.498	Stepchild (Yes = 1)	-3.574	0.503
Year (vs. 1998)			Children (Yes = 1)	-0.414	0.154
2000	-0.152	0.084	College (Yes = 1)	0.183	0.142
2002	0.019	0.092	Parent raised child	0.154	0.250
2004	0.072	0.106	Parent finan. help	-0.205	0.184
<i>Mother's characteristics</i>			<i>Family characteristics</i>		
Health (vs. Excellent)			Family size (vs. 1)		
Very good	-0.105	0.173	2	-1.052	0.181
Good	0.420	0.169	3	-1.538	0.187
Fair	0.701	0.173	4	-1.967	0.201
Poor	0.867	0.182	5–6	-2.508	0.207
Age (vs. 75–79)			7+	-2.521	0.224
70–74	-0.552	0.177	% Children		
80–84	0.482	0.096	Male	0.946	0.203
85–89	0.928	0.123	Married	-0.051	0.202
90+	1.213	0.156	Stepchild	0.940	0.478
Assets (dollars)			Have children	0.464	0.236
(vs. 100,000–249,000)			Attended college	-0.136	0.192
Negative	-0.336	0.258	Family got help (vs. No)		
0	0.004	0.151	Yes	0.595	0.187
<25,000	0.070	0.118	Missing	1.300	0.290
25,000–49,999	0.234	0.128			
50,000–99,999	0.171	0.111			
250,000+	-0.184	0.137			
Final illness	1.411	0.088			

Source: Results taken from Table 2 in J. Henretta et al., *J. Marriage & Family*, 73: 383–395, 2011. Reprinted with permission of J. Wiley & Sons.

The estimated variance components were 4.38 ($SE = 0.32$) for the $v_{i(j)}$ child random effects and 1.20 ($SE = 0.18$) for the u_j family random effects. As we'd expect, these reflect an especially strong within-child correlation in the repeated responses. Since $4.38/(4.38 + 1.20 + \pi^2/3) = 0.49$, variability among the children accounts for 49% of the total residual variance. Since $1.20/(4.38 + 1.20 + \pi^2/3) = 0.14$, family membership accounts for 14% of the total residual variance. This measure of variability among families also describes the degree to which siblings in a family are similar to each other, net of the explanatory variables in the model. The variability among families was more substantial in simpler models that did not include explanatory variables pertaining to the families such as mother's characteristics.

More complex models with interaction terms did not fit significantly better. In the other direction, the model shown in [Table 13.13](#) could be simplified by removing some nonsignificant factors or by treating the ordinal factors health, age, and family size in a quantitative manner and describing such effects by trends.

13.6 GLMM FITTING, INFERENCE, AND PREDICTION

Model fitting is rather complex for GLMMs, because the likelihood function does not have a closed form. Numerical methods for approximating it can be computationally intensive for models with multivariate random effects. In this section we outline the basic ideas of ML fitting. See Fahrmeir and Tutz (2001, Chap. 7) and McCulloch et al. (2008) for more details.

13.6.1 Marginal Likelihood and Maximum Likelihood Fitting

The GLMM is a two-stage model. At the first stage, conditional on the random effects $\{\mathbf{u}_i\}$, observations are assumed to follow a GLM. That is, all observations are independent, with y_{it} in cluster i having distribution in the exponential family with expected value μ_{it} linked to a linear predictor,

$$g(\mu_{it}) = \mathbf{x}_{it}^T \boldsymbol{\beta} + \mathbf{z}_{it}^T \mathbf{u}_i.$$

Then, $\mathbf{z}_{it}^T \mathbf{u}_i$ is a known offset. At the second stage, $\{\mathbf{u}_i\}$ are assumed independent from a $N(\mathbf{0}, \Sigma)$ distribution.

For a discrete variable, denote the vector of observations by \mathbf{y} and the vector of random effects by \mathbf{u} . Let $f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta})$ denote the conditional mass function of \mathbf{y} , given \mathbf{u} . Let $f(\mathbf{u}; \Sigma)$ denote the normal probability density function for \mathbf{u} . The likelihood function $\ell(\boldsymbol{\beta}, \Sigma; \mathbf{y})$ for a GLMM is the probability mass function $f(\mathbf{y}; \boldsymbol{\beta}, \Sigma)$ of \mathbf{y} , viewed as a function of $\boldsymbol{\beta}$ and Σ . This mass function refers to the marginal distribution of \mathbf{y} after integrating out the random effects,

$$(13.21) \quad \ell(\boldsymbol{\beta}, \Sigma; \mathbf{y}) = f(\mathbf{y}; \boldsymbol{\beta}, \Sigma) = \int f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}) f(\mathbf{u}; \Sigma) d\mathbf{u}.$$

It is often called a *marginal likelihood*. For example, the marginal likelihood function $\ell(\boldsymbol{\beta}, \sigma^2; \mathbf{y})$ for the logistic-normal random intercept model (13.6) (absorbing α into $\boldsymbol{\beta}$) is

$$\prod_i \left\{ \int_{-\infty}^{\infty} \prod_i \left[\frac{\exp(\mathbf{x}_{it}^T \boldsymbol{\beta} + u_i)}{1 + \exp(\mathbf{x}_{it}^T \boldsymbol{\beta} + u_i)} \right]^{y_{it}} \left[\frac{1}{1 + \exp(\mathbf{x}_{it}^T \boldsymbol{\beta} + u_i)} \right]^{1-y_{it}} f(u_i; \sigma^2) du_i \right\}.$$

Many methods can evaluate this numerically and maximize it as a function of $\boldsymbol{\beta}$ and Σ . It is an active area of research to develop improved methods. We next discuss a few of the most popular methods.

13.6.2 Gauss–Hermite Quadrature Methods for ML Fitting

The integral determining the likelihood function has dimension that depends on the random effects structure. When the dimension is small, as in the one-dimensional integral above, standard numerical integration methods can approximate the likelihood function.

Gauss–Hermite quadrature is a method for approximating the integral of a function $f(\cdot)$ multiplied by a normal density function. The approximation is a finite weighted sum that evaluates the function at certain points. In the univariate normal random effects case, the approximation has the form

$$\int_{-\infty}^{\infty} f(u) \exp(-u^2) du \approx \sum_{k=1}^q c_k f(s_k),$$

with *weights* $\{c_k\}$ and *quadrature points* $\{s_k\}$ that are tabulated. The approximation improves as q , the number of quadrature points, increases.

The approximated likelihood can be maximized with standard algorithms such as Newton–Raphson, yielding ML estimates $\hat{\beta}$ and $\hat{\Sigma}$. Inverting an approximation for the observed information matrix provides standard errors for the ML estimates. For complex models, second partial derivatives for the Hessian may be computed numerically rather than analytically. Adequate approximation usually requires larger q for standard errors than for $\hat{\beta}$. We recommend sequentially increasing q until the changes are negligible in both the estimates and standard errors.

When the function f to be integrated is not centered at 0, many of the quadrature points may fall outside the main region of integration. An adaptive version of Gauss–Hermite quadrature (Liu and Pierce 1994, Rabe-Hesketh et al. 2005) centers the quadrature points with respect to the mode of the function being integrated and scales them according to the estimated curvature at the mode. This improves efficiency, dramatically reducing the number of quadrature points needed to approximate the integrals effectively. Lesaffre and Spiessens (2001) showed comparisons and warned against using too few quadrature points.

13.6.3 Monte Carlo and EM Methods for ML Fitting

Multivariate forms of Gauss–Hermite quadrature handle multivariate, correlated random effects. Adequate approximation becomes more difficult, however, as the dimension of the integral increases much beyond the bivariate case. Then, Monte Carlo methods are more feasible computationally than numerical integration. Various Monte Carlo approaches are available [e.g., McCulloch et al. (2008, Chap. 14)], including Monte Carlo in combination with Newton–Raphson, Monte Carlo in combination with the EM algorithm, and simulation to estimate the likelihood directly. Here, we briefly describe a Monte Carlo EM (MCEM) algorithm.

The EM algorithm is a popular iterative method of finding ML estimates when data are missing or when filling in some “missing” data simplifies a likelihood function. Laird (2005) and Fan et al. (2010) gave useful reviews. In each cycle, an *E*-step takes an expectation over the missing data at working values of the parameters to approximate the likelihood function, and an *M*-step maximizes that function to generate new working values of the parameter estimates. With GLMMs, we regard the random effects \mathbf{u} as missing data. Then, $h\{\mathbf{y}, \mathbf{u}; \boldsymbol{\beta}, \boldsymbol{\Sigma}\} = f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta})f(\mathbf{u}; \boldsymbol{\Sigma})$ specifies the joint distribution of the complete data. The *E*-step in iteration r of the EM algorithm calculates, using Monte Carlo methods,

$$E[\log h(\mathbf{y}, \mathbf{u}; \boldsymbol{\beta}, \boldsymbol{\Sigma}) | \mathbf{y}; \boldsymbol{\beta}^{(r)}, \boldsymbol{\Sigma}^{(r)}].$$

The expectation refers to the distribution of $(\mathbf{u}|\mathbf{y})$ with parameter values equal to $\boldsymbol{\beta}^{(r)}$ and $\boldsymbol{\Sigma}^{(r)}$, the working estimates for iteration r . The distribution of $(\mathbf{u}|\mathbf{y})$ follows from those of $(\mathbf{y}|\mathbf{u})$ and \mathbf{u} in the GLMM via Bayes’ theorem. The *M*-step then maximizes (using MCMC or other Monte Carlo methods) the resulting function of $(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ to obtain $\boldsymbol{\beta}^{(r+1)}$ and $\boldsymbol{\Sigma}^{(r+1)}$. For details, including ways of choosing an appropriate Monte Carlo sample size, see Booth and Hobert (1999). Unfortunately, this method can also be impractical for large problems.

13.6.4 Laplace and Penalized Quasi-likelihood Approximations to ML

The Gauss–Hermite and Monte Carlo integration methods provide likelihood approximations such that resulting parameter estimates converge to the ML estimates as they are applied more finely, that is, as the number of quadrature points increases for Gauss–Hermite integration and as the Monte Carlo sample size increases in the MCEM method. This contrasts with other approximate methods that are simpler but do not yield ML estimates. These methods maximize an analytical approximation of the likelihood function.

Recall that the likelihood function (13.21) results from integrating out the random effects \mathbf{u} from the joint distribution of \mathbf{y} and \mathbf{u} . Using the exponential family representation of each component of that joint distribution, the integrand of (13.21) is an exponential function of \mathbf{u} . One approach approximates that function using a second-order Taylor series expansion of its exponent around a point $\tilde{\mathbf{u}}$ at which the first-order term equals 0. [That point $\tilde{\mathbf{u}} \approx E(\mathbf{u}|\mathbf{y})$.] The approximating function for the integrand is then exponential with quadratic exponent in $(\mathbf{u} - \tilde{\mathbf{u}})$ and has the form of a constant multiple of a multivariate normal density. Thus, its integral has closed form. This type of integral approximation is called a *Laplace approximation*. The approximation for integral (13.21) is then treated as a likelihood and maximized with respect to $\boldsymbol{\beta}$ and Σ .

For one such method (Breslow and Clayton 1993), the integral approximation yields a function approximating the log likelihood that has the form

$$q(\boldsymbol{\beta}, \mathbf{y}) = (1/2)\tilde{\mathbf{u}}^T \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{u}},$$

where $q(\boldsymbol{\beta}, \mathbf{y})$ resembles a quasi-log-likelihood function for the GLM conditional on $\mathbf{u} = \tilde{\mathbf{u}}$. Thus, the approximation results in a penalty for the quasi-log likelihood, with the penalty increasing as elements of $\tilde{\mathbf{u}}$ increase in absolute value. This approach is called *penalized quasi-likelihood* (PQL).

The calculations for maximizing the penalized quasi-likelihood use methods for linear mixed models with a normal response. This treats a linearization of the logit as a working response and entails iterative solution of sets of likelihood-like equations in $\boldsymbol{\beta}$ and \mathbf{u} . PQL methods do not require numerical or Monte Carlo integration and so are computationally simpler than ML methods. Unfortunately, they can perform poorly relative to ML (Breslow and Lin 1995). When true variance components are large, ordinarily PQL tends to produce variance component estimates with substantial negative bias. The PQL estimators also behave poorly when the response distribution is far from normal (e.g., binary).

Refinements have been and are being developed to lessen the bias of methods based on the Laplace approximation. These are useful, because precise ML inference using Gauss–Hermite quadrature or Monte Carlo methods is still impractical for large problems with GLMMs. See Zipunnikov and Booth (2012) for one such method and for relevant references.

13.6.5 Inference for GLMM Parameters

After fitting the model, inference about fixed effects proceeds in the usual way. For instance, likelihood-ratio tests can compare nested models. Asymptotics for GLMMs apply as the number of clusters increases, rather than as the numbers of observations within the clusters increase. Similarly, resampling methods such as the bootstrap using a large number of clusters should sample clusters rather than individual observations within clusters, to preserve the within-cluster dependence.

Inference about random effects (e.g., their variance components) is more complex. For instance, sometimes one model is a special case of another in which a variance component equals 0. The simpler model then falls on the boundary of the parameter space relative to the more complex model, so ordinary likelihood-based inference does not apply. For the most common situation, testing $H_0: \sigma^2 = 0$ against $H_a: \sigma^2 > 0$ for a model containing a random intercept, the null asymptotic distribution of the likelihood-ratio statistic is an equal mixture of χ_0^2 (i.e., degenerate at 0) and χ_1^2 random variables (Self and Liang 1987). The value of 0 occurs when $\hat{\sigma} = 0$, in which case the maximized likelihoods are identical under H_0 and H_a . When $\hat{\sigma} > 0$ and the observed test statistic equals t , the P -value for this large-sample test is $\frac{1}{2} P(\chi_1^2 > t)$, half the P -value that applies for χ_1^2 asymptotic tests. For testing more than one variance component, the mixture distribution is more complex (Molenberghs and Verbeke 2007).

13.6.6 Prediction Using Random Effects

We've used random effects in models to represent heterogeneity of certain characteristics, such as probabilities or odds ratios. Estimated effects of interest are often then linear combinations of fixed and random effects. For example, in the clinical trial comparing two treatments with random effects for centers (Section 13.3.5), we can predict the probability of success for each treatment in each center and odds ratios in those centers.

Given the data, the conditional distribution of $(\mathbf{u}|\mathbf{y})$ contains the information about the random effects \mathbf{u} . A prediction for \mathbf{u} is $E(\mathbf{u}|\mathbf{y})$, its *posterior mean* given the data. Calculation of $E(\mathbf{u}|\mathbf{y})$ itself requires numerical integration or Monte Carlo approximation. The expectation depends on β and Σ , so in practice we substitute $\hat{\beta}$ and $\hat{\Sigma}$ in the approximation. The standard error of the predictor of the random effect u_i is the standard deviation of the distribution of $(u_i|\mathbf{y})$. When we substitute $\hat{\beta}$ and $\hat{\Sigma}$ in $E(\mathbf{u}|\mathbf{y})$, however, the standard error does not account for the sampling variability in those estimates. Hence, the true standard error tends to be underestimated (Booth and Hobert 1998).

This approach to prediction using posterior means of random effects provides effect estimates that exhibit shrinkage relative to estimates using only data in the specific cluster. In this sense the results are similar to those using an *empirical Bayes* approach (Section 3.6.7, Efron and Morris 1975, Ten Have and Localio 1999). Shrinkage estimators can be far superior to sample values when the sample size for estimating each parameter is small, when there are many parameters to estimate, or when the true parameter values are roughly equal.

13.7 BAYESIAN MULTIVARIATE CATEGORICAL MODELING

With the Bayesian approach to GLMMs, the distinction between fixed and random effects need no longer occur, as every effect has a probability distribution. However, there is still the distinction between cluster-specific versus population-averaged effects according to whether the linear predictor contains a term for each cluster.

13.7.1 Marginal Homogeneity Analyses for Matched Pairs

For matched-pairs binary data without explanatory variables, Altham (1971) proposed Bayesian analyses. In the simplest case, she considered a model in which the probability of success is assumed the same for each subject at a given occasion. She showed that the classical exact P -value for testing the null hypothesis of marginal homogeneity (Section 11.1.5, using the binomial distribution) is also a Bayesian posterior probability for a Dirichlet prior distribution favoring H_0 .

Altham also used a model similar to (11.8) in which the probability varies by subject but the occasion effect is constant. She showed that the Bayesian evidence against the null hypothesis is weaker as the number of pairs giving the same response at both occasions increases, for fixed values of the numbers of pairs giving different responses at the two occasions. This differs from the frequentist result, using the McNemar test or conditional ML or random effects marginal ML, which does not depend on such pairs (e.g., Section 11.2.3). Consonni and La Rocca (2008) and Ghosh et al. (2000) showed related results.

13.7.2 Bayesian Approaches to Meta-analysis and Multicenter Trials

In Sections 13.3.5 and 13.3.6 we used random effects models to summarize heterogeneity in multicenter clinical trials and in meta-analyses. Comparable Bayesian analyses can use similar distributions for the random effects but also for the fixed effects.

Skene and Wakefield (1990) modeled multicenter binary-response studies with a logistic model that allows the treatment log odds ratio effect to vary among centers. They parameterized the model in terms of the logit for one group (identified as a placebo) and the log odds ratio comparing it to the second group. Conditional on those parameters, they assumed independent binomial distributions for the two groups in each center. For the logit and log odds ratio, they assumed exchangeability among centers, with a bivariate normal prior having the five hyperparameters unspecified but with their own second-stage prior. They treated the normal mean vector and covariance matrix as independent, with improper uniform priors for the means. They used various inverse Wishart prior distributions for the covariance, and conducted a sensitivity study to study the extent to which posterior inferences depended on that choice. The marginal posterior distributions of the mean and variance of the log odds ratio component then describe the difference between treatments and its heterogeneity among centers.

Skene and Wakefield also suggested forming a predictive distribution for the log odds ratio for a new center that was not part of the study but could be considered exchangeable with the ones in the study. This provides a way of assessing how the treatment would perform compared with placebo in a new center.

Warn et al. (2002) proposed Bayesian approaches to the meta-analysis issue discussed in Section 13.3.6 of allowing variability in the difference of proportions and relative risk. This addresses the difficulty of using normal distributions for parameters that have bounds on their values. Consider the model for binomial parameters whereby for study i ,

$$P(Y_{i1} = 1) = \pi_{i1}, \quad P(Y_{i2} = 1) = \pi_{i1} + \delta_i,$$

where $\delta_i = \pi_{i2} - \pi_{i1}$, assuming that $\{\delta_i\}$ have a $N(\delta, \tau^2)$ distribution. To reflect that π_{i1} and π_{i2} are constrained to $[0, 1]$, Warn et al. (2002) expressed $P(Y_{i2} = 1)$ as

$$P(Y_{i2} = 1) = \pi_{i1} + \min[\max(\delta_i, -\pi_{i1}), 1 - \pi_{i1}],$$

the increment to π_{i1} ensuring that π_{i2} falls in $[0, 1]$. For other prior distributions, Warn et al. (2002) suggested a uniform distribution over $(-1, 1)$ for δ and a uniform distribution over $(0, 2)$ for τ , which contains all the plausible values for τ . Possible priors for $\{\pi_{i1}\}$ included a beta distribution with uniform (1, 100) hyperpriors on the beta parameters, which are unimodal but relatively uninformative. This approach also has some awkward aspects. For instance, whenever δ_i is sampled outside the range $[-\pi_{i1}, 1 - \pi_{i1}]$, π_i is set to 0 or 1, resulting in spikes of probability at these values.

Casella and Moreno (2005) and Efron (1996) also proposed Bayesian methods for summarizing information from several 2×2 tables. Efron used empirical Bayesian methods to summarize odds ratios from 41 different trials of a surgical treatment for ulcers. His method permits selection from a wide class of priors in the exponential family.

13.7.3 Example: Bayesian Analyses for a Multicenter Trial

Skene and Wakefield illustrated their methodology with a Bayesian analysis of [Table 13.7](#), from the study to compare placebo with a drug for curing a fungal infection that we analyzed with GLMMs in Section 13.3.5. They noted that the data show evidence of a decrease in the treatment effect when the placebo success rate increases. Varying the Wishart second stage prior for the covariance of the normal prior for the placebo logit and the log odds ratio had an effect on the posterior distribution of the variance of the log odds ratio but little effect on the posterior distribution of its mean.

The posterior distribution for the mean of the log odds ratios had mean falling between 0.82 and 0.99 for the various priors, and standard deviation falling between 0.42 and 0.52. By contrast, the GLMM analysis with treatment \times center interaction estimated that the log odds ratios have a mean of 0.75, with standard error 0.32. The Bayesian analyses also reported posterior probabilities of a negative mean treatment effect (typically about 0.03), analogous to the one-sided P -value in the GLMM analysis (which was 0.01).

With the various priors, the mean of the posterior distribution of the variance of the log odds ratio varied between 0.49 and 1.03, reflecting potentially considerable heterogeneity among centers in the true effect. For a new center for the study summarized by [Table 13.7](#), the predictive density for the log odds ratio was considerably wider than the posterior density for the mean log odds ratio. With a typical prior, it gave probability 0.19 of a negative value, the relatively large value reflecting heterogeneity among centers.

13.7.4 Bayesian GLMMs and Marginal Models

Bayesian methods have been used to approximate ML fitting of GLMMs. Use of a flat prior distribution yields a posterior density that is a constant multiple of the likelihood function. Then, Markov chain Monte Carlo (MCMC) methods for approximating intractable posterior distributions can approximate the likelihood function (Zeger and Karim 1991). An approximation for the mode of the posterior distribution approximates the ML estimate.

A danger is that improper prior distributions have improper posteriors for many models for categorical data (Natarajan and McCulloch 1995). In using MCMC, we may fail to realize that the posterior is improper. It is safer to use a proper but relatively diffuse prior. With a multivariate normal prior, we could also use a hierarchical approach in which the covariance matrix has an inverse Wishart distribution. However, the posterior mode need not be close to the ML estimate, and Markov chains may converge slowly.

Of course, Bayesian methods can be used not only to approximate frequentist results but also as a standard approach for those who prefer the Bayesian paradigm, whether it be for estimating population-averaged or cluster-specific effects. For example, Daniels and Gatsonis (1999) used multilevel GLMs to analyze geographic and temporal trends with clustered longitudinal binary data. This built on hierarchical modeling ideas introduced by Wong and Mason (1985).

We've seen that logistic regression does not extend easily to the modeling of multivariate categorical responses, because of a lack of a simple logistic analog of the multivariate normal. However, O'Brien and Dunson (2004) formulated a multivariate logistic distribution incorporating correlation parameters and having marginal logistic distributions. They used this in a Bayesian analysis of marginal logistic regression models, showing that proper posterior distributions typically exist even with an improper uniform prior for the regression parameters.

For modeling multivariate correlated ordinal responses, Chib and Greenberg (1998) used a multivariate probit model. A multivariate normal latent random vector with cutpoints along the real line defines the categories of the observed discrete variables. The correlation among the categorical responses is induced through the covariance matrix for the underlying latent variables. Webb and Forster (2008) parameterized the model in such a way that conditional posterior distributions are standard and easily simulated. They focused on model determination through comparing posterior marginal probabilities of the model given the data (integrating out the parameters).

Chen and Shao (1999) briefly reviewed other Bayesian approaches to handling such data. They employed a scale mixture of multivariate normal links, a class of models that includes the multivariate probit, t link, and logit. Chen and Shao offered both a noninformative and an informative prior and gave conditions ensuring that the posterior is proper. Note 13.13 lists other references dealing with Bayesian multivariate categorical data analysis.

NOTES

Section 13.1: Random Effects Modeling of Clustered Categorical Data

13.1 Rasch, clustered binary references: For further discussion of the Rasch model and ways of estimating its parameters, see Andersen (1980, Sec. 6.4) and Fischer and Molenaar (1995). Haberman (1977b) showed that ML estimators can achieve consistency when both n and T grow at suitable rates. Early work on GLMMs for a categorical response includes Anderson and Aitkin (1985), Bartholomew (1980), Bock and Aitkin (1981), Chamberlain (1980), Gilmour et al. (1985), Pierce and Sands (1975), and Stiratelli et al. (1984). Caffo and Griswold (2006) and Caffo et al. (2007) discussed probit random effects models and related models using the t link. Hedeker and Gibbons (2006), Molenberghs and Verbeke (2005), Neuhaus (1992), and Pendergast et al. (1996) surveyed methods for clustered binary data, including GLMMs and marginal models. McCullagh (2008) argued that most natural sampling schemes involving binary random effects models are biased, an implication being that the effects for such models and corresponding marginal models are not necessarily the relevant effects.

13.2 Conditional ML versus random effects: In models with covariates, Neuhaus and Lesperance (1996) noted that conditional ML may lose efficiency compared with the random effects approach when cluster sizes are small and covariates have strong positive within-cluster correlation. As that correlation approaches +1, the covariate effect resembles a between-cluster one, which the conditional ML approach cannot estimate. The matched-pairs case referred to in Section 13.1.2 in which the conditional ML estimate equals the random effects estimate has within-cluster covariate correlation = -1, as depending on the order of viewing the observations, x_t changes from 0 to 1 or from 1 to 0; then, no efficiency loss occurs.

13.3 Nonnormal random effects: Alternatives to the normal random effects distribution are conjugate random effects, a mixture of normals, and a combination of conjugate random effects and normal random effects. See Caffo et al. (2007), Molenberghs et al. (2010), and Lee et al. (2006). Wang and Louis (2003) and Parzen et al. (2011) showed that when the random effects in a logistic model have a certain scale mixture of normal distributions, the marginal model also has logistic form.

Section 13.3: Examples of Random Effects Models for Binary Data

13.4 Capture–recapture, heterogeneity: For other analyses permitting heterogeneous odds ratios in several 2×2 tables, see Casella and Moreno (2005), Efron (1996), Liu and Pierce (1993), and Skene and Wakefield (1990). For further discussion of capture–recapture modeling, see Bishop et al. (1975, Chap. 6), Chao et al. (2001), Cormack (1989), Coull and Agresti (1999), Darroch et al. (1993), Fienberg et al. (1999), Hook and Regal (1995), Pledger et al. (2010), Royle et al. (2007), and the many references in these articles. Similarities exist between this problem and the related problem of estimating the binomial index n when observing independent $\text{bin}(n, \pi)$ counts with unknown n and π ; see DasGupta and Rubin (2005) and Grevstad (2006) and references in those articles. Relatively flat log likelihoods also occur with other models that permit capture heterogeneity (Burnham and Overton 1978), such as a beta-binomial model.

13.5 Meta-analysis: For alternative random effects approaches to meta-analysis, see Burr and Doss (2005), Efron (1996), Emerson et al. (1993), Rücker et al. (2009), Shuster (2010), Stijnen et al. (2010), and Tian et al. (2009).

13.6 Ecological inference: King (1997) used random effects models as part of a solution for analyzing aggregated categorical data, the problem of *ecological inference*. Chambers and Steel (2001) discussed early work by Leo Goodman on this problem and proposed a simpler

semiparametric approach. See also Wakefield (2004).

13.7 Joint response models: For longitudinal bivariate binary responses, Ten Have and Morabia (1999) simultaneously modeled bivariate log odds ratios and univariate logits. Multivariate responses sometimes have both continuous and categorical components. For random effects modeling of such data, see Catalano and Ryan (1992), Gueorguieva and Agresti (2001), and Molenberghs and Verbeke (2005, Chap. 24). See Gueorguieva (2001) for a multivariate generalization.

13.8 Spatial data: For examples of random effects models for spatial categorical response data, see Banerjee et al. (2004), Heagerty and Lele (1998), Hoeting et al. (2000), Kneib and Fahrmeir (2006), and Miller and Franklin (2002).

Section 13.4: Random Effects Models for Multinomial Data

13.9 Ordinal response: The same predictor structure as in (13.18) holds with other links for which a common effect for each logit is plausible, such as adjacent-categories logits (Hartzel et al. 2001a,b). With the complementary log–log link, the likelihood function has closed form with a log gamma random effects distribution (Crouchley 1995 Ten Have 1996). Agresti and Natarajan (2001), Hedeker and Gibbons (1994; 2006, Chap. 10), Hartzel et al. (2001a,b), Hedeker (2008), and Tutz and Hennevogl (1996) presented ordinal response models with random effects.

13.10 Nominal response: For multinomial extensions of the Rasch model, see Andersen (1980, pp. 272–284; 1995) and Conaway (1989). Daniels and Gatsonis (1997), Hartzel et al. (2001b), Hedeker (2008), and Hedeker and Gibbons (2006, Chap. 11) presented nominal-response models with random effects. For discrete choice models (Section 8.5) with random effects, see Chen and Kuo (2001), McFadden and Train (2000), Natarajan et al. (2000), and Train (2009).

Section 13.5: Multilevel Models

13.11 Multilevel references: Early work on multilevel modeling for categorical data includes Aitkin et al. (1981), Anderson and Aitkin (1985), and Wong and Mason (1985). For later work, see Browne et al. (2005), Carlin et al. (2001), Daniels and Gatsonis (1997, 1999), Gelman and Hill (2006, Ch. 14, 15), Gibbons and Hedeker (1997), Goldstein (2010), Guo and Zhao (2000), Heagerty and Zeger (2000) for a marginal approach, Hedeker (2008) for a survey for nominal and ordinal data, Longford (1993), Skrondal and Rabe-Hesketh (2003, 2004), and Vermunt (2003) for multilevel latent class models, and Yang et al. (2000).

Section 13.6: GLMM Fitting, Inference, and Prediction

13.12 Marginally specified model: A GLMM determines the marginal relationship (averaged over random effects) between the mean response and explanatory variables. Conversely, Heagerty (1999) noted that a marginal model for the mean implicitly determines the form of the fixed portion of the linear predictor in a random effects model. The GLMM (13.1) has linear predictor, $\mathbf{x}_{it}^T \boldsymbol{\beta} + \mathbf{z}_{it}^T \mathbf{u}_i$. A more general form $\Delta_{it} + \mathbf{z}_{it}^T \mathbf{u}_i$ corresponds to a particular marginal model. Here, Δ_{it} is a function of the marginal linear predictor and the random effects distribution. It is implicitly defined by the integral equation that links the marginal and conditional means. Caffo et al. (2007) gave related discussion, and Swihart et al. (2012) discussed equivalent copula models.

Section 13.7: Bayesian Multivariate Categorical Modeling

13.13 Bayes multivariate: Dey et al. (2000) edited a collection of articles that provided Bayesian analyses for GLMs, often in a multivariate setting. For instance, in that volume Gelfand and Ghosh surveyed the subject, Albert and Ghosh reviewed item-response modeling, Chib modeled correlated binary data, Chen and Dey modeled correlated ordinal data, and Landrum

and Normand gave a case study using Bayesian ordinal probit and logit models.

EXERCISES

Applications

13.1 Refer to the heaven/hell matched-pairs data of [Table 11.14](#) and Exercise 11.8.

- a. Fit the random intercept model [\(13.3\)](#). Interpret $\hat{\beta}$.
- b. Compare $\hat{\beta}$ and its SE for this approach to the conditional ML approach.
- c. Refer to the two logistic models used in Exercise 11.8. Explain why the population-averaged and subject-specific effects differ so much for these data.

13.2 Refer to [Table 4.11](#) on the three-point shooting of Ray Allen. In game i , suppose that y_i = number made out of n_i attempts is a $\text{bin}(n_i, \pi_i)$ variate and $\{y_i\}$ are independent.

- a. Fit the model, $\text{logit}(\pi_i) = \alpha$. Find and interpret $\hat{\pi}_i$. Does the model appear to fit adequately?
- b. Fit the model, $\text{logit}(\pi_i) = \alpha + u_i$, where $\{u_i\}$ are independent $N(0, \sigma^2)$. Use $\hat{\alpha}$ and $\hat{\sigma}$ to summarize Allen's shooting. Is there evidence that this model fits better than the one in part (a)?

13.3 For [Table 9.3](#), let $y_{it} = 1$ when subject i used substance t . [Table 13.14](#) shows output for the logistic-normal model

Table 13.14 Output for Exercise 13.3 on Cigarettes, Alcohol, and Marijuana Use

Subjects	2276	Parameter	Estimate	Std Error	t Value
Max Obs Per Subject	3	beta1	4.2227	0.1824	23.15
Parameters	4	beta2	1.6209	0.1207	13.43
Quadrature Points	200	beta3	-0.7751	0.1061	-7.31
Log Likelihood	-3311	sigma	3.5496	0.1627	21.82

$$\text{logit}[P(Y_{it} = 1|u_i)] = \beta_t + u_i.$$

Interpret $\hat{\sigma}$ and the effect that compares use of cigarettes ($t = 2$) and marijuana ($t = 3$). How is the focus different from that for the loglinear model (*AC*, *AM*, *CM*) used in Section 9.2.4? If $\hat{\sigma} = 0$, which loglinear model would have the same fit as this GLMM?

13.4 For the student survey data in [Table 10.1](#), (a) analyze using GLMMs, and (b) compare results and interpretations to those with marginal models in Exercise 12.2.

13.5 Consider model [\(13.11\)](#) for the attitudes toward abortion data in [Table 13.3](#).

- a. Fit the model. If your software uses Gauss–Hermite numerical integration, report $\{\hat{\beta}_t\}$ and their standard errors for 5, 25, 100, and 500 quadrature points, and comment on convergence.
- b. Under the constraint $\sigma = 0$, explain why the fit is the same as (i) an ordinary logistic model treating the three responses for each subject as if they were independent responses for three separate subjects, (ii) an ordinary loglinear model (GS_1, GS_2, GS_3) of mutual independence of responses in the three situations (S_1, S_2, S_3), given $G = \text{gender}$.
- c. Fit one of the models in (b). Interpret, and explain why $\{\hat{\beta}_t - \hat{\beta}_u\}$ are quite different from the estimates in Section 13.3.2 for the model allowing $\sigma > 0$.

13.6 Consider the crossover study in [Table 12.9](#) (Exercise 12.6).

- a. Fit the model

$$(13.22) \text{logit}[P(Y_{i(k)t} = 1|u_{i(k)})] = \alpha_k + \beta_t + u_{i(k)},$$

where $\{u_{i(k)}\}$ are independent $N(0, \sigma^2)$. Interpret $\{\hat{\beta}_t\}$ and $\hat{\sigma}$.

- b. We can also add period or carryover effects. Add two period effects to model [\(13.22\)](#) (e.g., the first-period-effect parameter adds to the model when $t = A$ and $k = 1, 2$, $t = B$ and $k = 3, 4$, and $t = C$ and $k = 5, 6$). Check whether the fit improves. Interpret.
- c. For the model in (a), compare estimates of $\beta_B - \beta_A$ and $\beta_C - \beta_A$ and SE values to those using (i) a marginal model, and (ii) conditional logistic regression, treating subject terms in model [\(13.22\)](#) as fixed effects.

13.7 For [Table 6.6](#) on admissions decisions for graduate school applicants, let $y_{ig} = 1$ when a subject in department i of gender g ($1 = \text{females}$, $0 = \text{males}$) is admitted.

- a. For the fixed effects model, $\text{logit}[P(Y_{ig} = 1)] = \alpha + \beta_g + \beta_i^D$, $\hat{\beta} = 0.173$ ($SE = 0.112$). The corresponding model [\(13.14\)](#) in which departments are a normal random effect has $\hat{\beta} = 0.163$ ($SE = 0.111$). Interpret these.
- b. The model of form [\(13.14\)](#) allowing the gender effect to vary by department has $\hat{\beta} = 0.176$ ($SE = 0.132$), with $\hat{\sigma}_b = 0.20$. Interpret. Explain why the standard error of $\hat{\beta}$ is larger than with the other analyses.
- c. The sample conditional odds ratios between gender and whether admitted vary between 0 and ∞ . By contrast, predicted odds ratios for the interaction random effects model do not vary much. Explain why.

13.8 For the clinical trial in [Table 6.11](#), let $\pi_{it} = P(Y_{it} = 1|u_i)$ denote the probability of success for treatment t in center i .

- a. The random intercept model [\(13.13\)](#) has $\hat{\beta} = 1.52$ ($SE = 0.70$) and $\hat{\sigma} = 1.9$. Interpret.
- b. From Section 6.5.2, the fixed effects analog of this model (replacing $\alpha + u_i$ by α_i) has $\hat{\alpha}_1 = \hat{\alpha}_3 = -\infty$, corresponding to $\hat{\pi}_{1t} = \hat{\pi}_{3t} = 0$ for each treatment. By contrast, the random effects model has $\hat{\alpha} + \hat{u}_1 = -3.78$ (using NLMIXED in SAS) and $\hat{\pi}_{11} = 0.047$ and $\hat{\pi}_{12} = 0.011$ in center 1. Explain how this model can have $\hat{\pi}_{it} > 0$ in centers having no successes.

13.9 For the subject-specific model in Section 13.3.3 for the depression study, verify that the estimated difference in time effect slopes between the new and standard drugs for treating depression are (a) 1.018 ($SE = 0.192$) with the GLMM approach, and (b) 1.156 ($SE = 0.222$) with conditional ML.

13.10 For marginal model [\(11.16\)](#) for [Table 11.6](#) on premarital and extramarital sex, [Table 13.15](#) shows results of fitting a corresponding random intercept model. Interpret $\hat{\beta}$. Why is the estimate so different from β in Section 11.3.4 for the marginal model?

Table 13.15 Output for Exercise 13.10 on GSS Items

Subjects	1337	Parameter	Estimate	Std Error	t Value
Max Obs Per Subject	2	inter1	-1.9702	0.1164	-16.93
Parameters	5	inter2	0.9840	0.0659	14.93
Quadrature Points	100	inter3	1.1737	0.0761	15.43
-2 Log Likelihood	5000.5	beta	4.2387	0.1975	21.46
		sigma	1.8612	0.1422	13.09

13.11 Landis and Koch (1977) showed ratings by seven pathologists who separately classified 118 slides regarding the presence and extent of carcinoma of the uterine cervix, using a five-point ordinal scale. ([Table 14.1](#) is a collapsing of their table that combines the first two categories and the last three categories.) For slide i with rater t , the model

$$\text{logit}[P(Y_{it} \leq j|u_i)] = \alpha_j + \beta_t + u_i$$

fitted (with $\beta_7 = 0$) assuming that $\{u_i\}$ are independent $N(0, \sigma^2)$ has $\hat{\beta}_6 = 2.907$ ($SE = 0.344$) and $\hat{\sigma} = 3.8$. The corresponding marginal model, fitted using independence working correlations, has GEE estimate $\hat{\beta}_6 = 1.252$ ($SE = 0.161$). Interpret $\hat{\beta}_6$ for each model. Explain why $\hat{\beta}_6$ for the GLMM is much larger in absolute value. Discuss the differences in assumptions and interpretations for the two models.

13.12 From Section 11.4.2, the ML estimates of main effects in the quasi-symmetry model relate to conditional ML estimates for a subject-specific model using baseline-category logits. For the migration data in [Table 11.7](#), $\hat{\lambda}_1^X - \hat{\lambda}_1^Y = 1.74$ when constraints set $\hat{\lambda}_4^X = \hat{\lambda}_4^Y = 0$. For a given subject, the estimated odds of living in the Northeast instead of the West at age 16 were $\exp(1.74) = 5.70$ times the odds in 2010. Explain why the corresponding population-averaged odds ratio estimate is $[(370/341)/(291/391)] = 1.46$, and explain how the estimates can differ so much.

13.13 Refer to Section 13.3.8 on boys' attitudes toward the leading crowd. [Table 13.16](#) shows results for a sample of schoolgirls. Fit model [\(13.16\)](#) and interpret. Summarize the estimated variability and correlation of random effects.

Table 13.16 Data for Exercise 13.13 on Girls and Leading Crowd

(M, A) for First Interview	(M, A) for Second Interview ^a			
	(Yes, Positive)	(Yes, Negative)	(No, Positive)	(No, Negative)
Yes, positive	484	93	107	32
Yes, negative	112	110	30	46
No, positive	129	40	768	321
No, negative	74	75	303	536

^a M , membership; A , attitude.

Source: J. S. Coleman, *Introduction to Mathematical Sociology*. London: Free Press of Glencoe, 1964, p. 168.

13.14 Generalize model [\(13.16\)](#) to apply simultaneously to [Table 13.8](#) for boys and [Table 13.16](#) for girls, using a gender main effect but the same membership effect and the same attitude effect for each gender. Fit the model. Use the maximized log likelihood to compare with a more general model having different membership effects and different attitude effects for each gender. Interpret.

13.15 [Table 13.17](#) reports results from a study to estimate the number N of people infected during a 1995 hepatitis A outbreak in Taiwan. The 271 observed cases were reported from records based on a serum test taken by the Institute of Preventive Medicine of Taiwan (P), records reported by the National Quarantine Service (Q), and records based on questionnaires administered by epidemiologists (E).

Table 13.17 Data for Exercise 13.15 on Hepatitis Infections

P Q E	Observed Count	Logistic-Normal ML Fit
0 0 0	—	(487, ∞)
0 0 1	63	61.0
0 1 0	55	58.0
0 1 1	18	17.0
1 0 0	69	68.0
1 0 1	17	20.0
1 1 0	21	19.0
1 1 1	28	28.0

Source: Data from Chao et al. (2001).

- a. Using the model of mutual independence with P, Q, and E, find \hat{N} and a 95% profile likelihood interval for N .
- b. The random effects model of Section 13.3.4 has fit shown in [Table 13.17](#), for which $\hat{\sigma} = 2.9$. The log likelihood is relatively flat, and $\hat{N} = 4551$ with a 95% profile likelihood interval of $(758, \infty)$ (Coull and Agresti 1999). Since the interval in part (a) is much narrower, is it necessarily more reliable? Explain.

13.16 Analyze the crossover data of [Table 11.22](#) using a random effects model. Interpret.

13.17 The analyses in Section 13.3.5 describing heterogeneity in multicenter clinical trials extend to ordinal responses. Using random effects models, analyze the $2 \times 3 \times 8$ table in Hartzel et al. (2001a), shown also at the text website.

13.18 Exercises 6.18 and 6.19 referred to published meta-analyses. For one of these, conduct a meta-analysis that uses methods of this chapter. Interpret.

13.19 For the example in Section 13.4.6, in interpreting effects, Hedeker (2008) reported sample proportions in each response category at each time for each group. He noted that over time, (a) there was a general decrease in street living and an increase in independent living for both groups, (b) the increase in independent living occurs sooner for the certificate group than the control group, (c) regarding community living, this increases for the control group and decreases for the certificate group. Explain how the estimates in [Table 13.12](#) suggest these interpretations.

13.20 Refer to Exercise 12.19 and the data for a clinical trial for toenail infection.

a. Fit a logistic-normal random intercept model for the binary endpoint. Discuss how the treatment effect estimate at baseline and its SE depend on the fitting method you use (e.g., on the number of quadrature points).

b. Compare results to those of a marginal model analysis for the data.

13.21 Analyze [Table 12.8](#) with age and maternal smoking as predictors using a (a) logistic-normal model, (b) marginal model, and (c) transitional model. Explain how the interpretation of the maternal smoking effect differs for the three approaches.

13.22 For Exercise 6.29 about a meta-analysis on the effect of rosiglitazone on myocardial infarction, conduct a fully Bayesian analysis. Justify the choice of priors. Compare results and interpretations to the fixed effects analysis.

Theory and Methods

13.23 For the voting example in Section 13.3.1, using supplementary information improves predictions. Let q_i denote the true proportion of votes for Kerry (the Democratic candidate) in state i in the 2004 election. Consider the model

$$\text{logit}[P(Y_{it} = 1|u_i)] = \text{logit}(q_i) + \alpha + u_i,$$

where $\{q_i\}$ are known and $\{u_i\}$ are independent $N(0, \sigma^2)$. When $\hat{\alpha} = 0$, show $\hat{q}_i = q_i \exp(\hat{\alpha})/[1 - q_i + q_i \exp(\hat{\alpha})]$. Compared to $\{q_i\}$, explain how \hat{q}_i then shifts up or down depending on how the overall Democratic vote compares in the current poll to the previous election (i.e., depending on $\hat{\alpha}$). When also $\hat{\alpha} = 0$, show $\hat{q}_i = q_i$.

13.24 For a binary response, consider the random effects model

$$\text{logit}[P(Y_{it} = 1|u_i)] = \alpha + \beta_t + u_i, \quad t = 1, \dots, T,$$

where $\{u_i\}$ are independent $N(0, \sigma^2)$, and the marginal model

$$\text{logit}[P(Y_t = 1)] = \alpha + \beta_t^*, \quad t = 1, \dots, T.$$

For identifiability, $\beta_T = \beta_T^* = 0$. Explain why all $\beta_t = 0$ implies that all $\beta_t^* = 0$. Is the converse true?

13.25 The GLMM for binary data using probit link function is

$$\Phi^{-1}[P(Y_{it} = 1|u_i)] = \mathbf{x}_{it}^T \boldsymbol{\beta} + \mathbf{z}_{it}^T \mathbf{u}_i,$$

where Φ is the $N(0, 1)$ cdf and \mathbf{u}_i has $N(\mathbf{0}, \boldsymbol{\Sigma})$ pdf, $f(\mathbf{u}_i; \boldsymbol{\Sigma})$.

a. Show that the marginal mean is

$$P(Y_t = 1) = \int P(Z - \mathbf{z}_{it}^T \mathbf{u}_i \leq \mathbf{x}_{it}^T \boldsymbol{\beta}) f(\mathbf{u}_i; \boldsymbol{\Sigma}) d\mathbf{u}_i,$$

where Z is a standard normal variate that is independent of \mathbf{u}_i .

b. Since $Z - \mathbf{z}_{it}^T \mathbf{u}_i$ has a $N(0, 1 + \mathbf{z}_{it}^T \boldsymbol{\Sigma} \mathbf{z}_{it})$ distribution, deduce that

$$\Phi^{-1}[P(Y_t = 1)] = \mathbf{x}_{it}^T \boldsymbol{\beta} [1 + \mathbf{z}_{it}^T \boldsymbol{\Sigma} \mathbf{z}_{it}]^{-1/2}.$$

Hence, the marginal model is a probit model with attenuated effect. In the univariate random intercept case, show that the marginal effect equals that from the GLMM divided by $\sqrt{1 + \sigma^2}$.

13.26 In the Rasch model, $\text{logit}[P(Y_{it} = 1)] = \alpha_i + \beta_t$, treat α_i as a fixed effect.

a. Assuming independence of responses for different subjects and for different observations on the same subject, show that the log likelihood is

$$\sum_i \sum_t \alpha_i y_{it} + \sum_i \sum_t \beta_t y_{it} - \sum_i \sum_t \log[1 + \exp(\alpha_i + \beta_t)].$$

b. Show that the likelihood equations are $y_{+t} = \sum_i P(Y_{it} = 1)$ and $y_{i+} = \sum_t P(Y_{it} = 1)$ for all i and t . Explain why conditioning on $\{y_{i+}\}$ yields a distribution that does not depend on $\{\alpha_i\}$.

13.27 Consider the matched-pairs random effects model [\(13.3\)](#). For given β_0 , let δ_0 be such that $\mu_{12} = n_{12} + \delta_0$ and $\mu_{21} = n_{21} - \delta_0$ satisfies $\log(\mu_{21}/\mu_{12}) = \beta_0$. Suppose $\{\mu_{ij}\}$ has nonnegative log odds ratio. Explain why:

- a. This is the fit of the model assuming $\beta = \beta_0$.
- b. The likelihood-ratio statistic for testing $H_0: \beta = \beta_0$ in this model equals

$$2 \left(n_{12} \log \frac{n_{12}}{n_{12} + \delta_0} + n_{21} \log \frac{n_{21}}{n_{21} - \delta_0} \right).$$

- c. The likelihood-ratio test of $H_0: \beta = 0$ is the test of symmetry.

13.28 Explain why the logistic-normal model is not helpful for capture–recapture experiments with only two captures.

13.29 In recent U.S. Presidential elections, in each state more wealthy voters tend to be more likely to vote Republican, yet states that are wealthier in an aggregate sense are more likely to go Democrat for the electoral college. Sketch a plot that illustrates how this instance of Simpson’s paradox could occur. Specify a GLMM with random effects for states that could be used to analyze data for a sample of voters using their state of residence, their household income, and their vote in an election. Explain how the model could be generalized to allow the income effect to vary by state. [For details, see Gelman and Hill (2007, Sec. 14.2).]

13.30 Summarize advantages and disadvantages of using a GLMM approach compared with a marginal model approach. Describe conditions under which parameter estimators are consistent for (a) marginal models using GEE, (b) marginal models using ML, and (c) GLMMs using ML.

¹Although see Note 13.12 for an implicit connection.

²The data and a SAS file for these analyses are at tigger.uic.edu/~hederker/long.html.

CHAPTER 14

Other Mixture Models for Discrete Data

In Chapters 11 through 13 we introduced methods for observations that are correlated due to repeated measurement and other forms of clustering. The generalized linear mixed models (GLMMs) of Chapter 13 assume normal random effects. They describe heterogeneity by replacing the linear predictor by a normally distributed mixture of linear predictors. In this chapter we present GLMM-type models that, except for one case, use nonnormal mixture distributions.

In Section 14.1 we present *latent class models*. These treat a contingency table as a finite mixture of unobserved tables generated under a conditional independence structure at categories of a latent variable. In Section 14.2 we present a related nonparametric approach to fitting GLMMs that uses an unspecified discrete quantitative distribution for the random effects distribution.

In Section 14.3 we present models for clustered binomial responses that use the beta distribution to describe heterogeneity of binomial parameters. The resulting *beta-binomial distribution* has variance function for which quasi-likelihood methods are also available. In Section 14.4 we model count responses using the gamma distribution to describe heterogeneity of Poisson parameters. The resulting *negative binomial* regression model corresponds to a Poisson GLMM having a log-gamma distributed random effect. It is an alternative to the GLMM for Poisson responses with normal random effects, a model presented in Section 14.5.

14.1 LATENT CLASS MODELS

Ordinary GLMMs create a mixture of linear predictor values using a latent variable, the unobserved random effect vector, that is assumed to have a normal distribution. By contrast, latent class models use a mixture distribution that is qualitative rather than quantitative. The basic model assumes existence of a latent categorical variable such that the observed response variables are conditionally independent, given that variable.

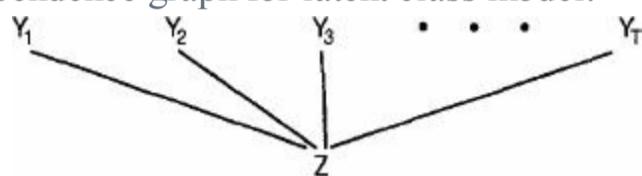
14.1.1 Independence Given a Latent Categorical Variable

For categorical response variables (Y_1, Y_2, \dots, Y_T) , the latent class model assumes a latent categorical variable Z such that for each possible sequence of response outcomes (y_1, \dots, y_T) and each category z of Z ,

$$P(Y_1 = y_1, \dots, Y_T = y_T | Z = z) = P(Y_1 = y_1 | Z = z) \cdots P(Y_T = y_T | Z = z).$$

The model was introduced by Lazarsfeld in 1950 and described by Lazarsfeld and Henry (1968). [Figure 14.1](#) shows the conditional independence graph for the model. A latent class model summarizes probabilities of classification $P(Z = z)$ in the latent classes as well as conditional probabilities $P(Y_t = y_t | Z = z)$ of outcomes for each Y_t within each latent class. These are the model parameters. The model is an analog for categorical responses and latent variables of the factor analysis model with a common factor for multivariate normal responses.

[Figure 14.1](#) Conditional independence graph for latent class model.



The latent class model is sometimes plausible when the observed variables are several indicators of some concept, such as prejudice, religiosity, or opinion about an issue. An example is [Table 13.3](#), in which subjects gave their opinions about whether abortion should be legal in various situations. Perhaps an underlying latent variable describes one's basic attitude toward legalized abortion, such that given the value of that latent variable, responses on the observed variables are conditionally independent. For instance, there may be three latent classes: one for those who always oppose legalized abortion regardless of the situation, one for those who always support it, and one for those whose response depends on the situation.

The T -dimensional contingency table cross-classifying (Y_1, \dots, Y_T) is observed. The $(T + 1)$ -dimensional table that cross-classifies it with the latent variable is an unobserved table. Denote the number of categories of each Y_t by I and the number of latent classes of Z by q . For the observed table, let $\pi_{y_1, \dots, y_T} = P(Y_1 = y_1, \dots, Y_T = y_T)$. The model assumes a multinomial distribution over its I^T cells. Each cell probability satisfies

$$\pi_{y_1, \dots, y_T} = \sum_{z=1}^q P(Y_1 = y_1, \dots, Y_T = y_T | Z = z) P(Z = z).$$

The conditional independence factorization for the latent class model states that

$$(14.1) \quad \pi_{y_1, \dots, y_T} = \sum_{z=1}^q \left[\prod_{t=1}^T P(Y_t = y_t | Z = z) \right] P(Z = z).$$

This is a nonlinear model for the I^T multinomial probabilities.

The latent class model implies that the loglinear model symbolized by $(Y_1 Z, Y_2 Z, \dots, Y_T Z)$ holds for the unobserved table. The model makes no assumption about the $\{Y_t Z\}$ associations but assumes that the $\{Y_t\}$ are mutually independent within each category of Z .

More generally, the latent variable can be multivariate. For example, for the membership and attitude toward the “leading crowd” data that we analyzed in Section 13.3.8 with model [\(13.16\)](#) with correlated normal random effects, Goodman (1974) used an analogous model with two associated binary latent variables.

14.1.2 Fitting Latent Class Models

Denote the counts in the observed table by $\{n_{y_1}, \dots, y_T\}$. For the I^T cells in that table, the kernel of the multinomial log likelihood is the sum over these cells,

$$(14.2) \sum n_{y_1, \dots, y_T} \log \pi_{y_1, \dots, y_T}.$$

Substituting (14.1), we can maximize this with respect to the model parameters $\{P(Y_t = y_t | Z = z)\}$ and $P(Z = z)\}$ using the EM algorithm (Goodman 1974) or the Newton–Raphson algorithm (Haberman 1979, Chap. 10).

The EM algorithm has two steps in each iteration. The E (expectation) step in iteration s calculates pseudo-counts $\{n_{y_1, \dots, y_T, z}^{(s)}\}$ for the unobserved table using $\{n_{y_1, \dots, y_T}\}$ and a working conditional distribution for $(Z|Y_1, \dots, Y_T)$ described shortly. The M (maximization) step treats $\{n_{y_1, \dots, y_T, z}^{(s)}\}$ as data and maximizes the pseudo-likelihood, fitting the loglinear model $(Y_1Z, Y_2Z, \dots, Y_TZ)$. The fit $\{\mu_{y_1, \dots, y_T, z}^{(s)}\}$ of that model in the unobserved table then determines the new working conditional distribution of $(Z|Y_1, \dots, Y_T)$ to apply to $\{n_{y_1, \dots, y_T}\}$ for the E -step of the next iteration. This allocates the observed data to pseudo-counts in the unobserved cells in proportion to this fit, using

$$n_{y_1, \dots, y_T, z}^{(s+1)} = n_{y_1, \dots, y_T} \frac{\mu_{y_1, \dots, y_T, z}^{(s)}}{\sum_{k=1}^q \mu_{y_1, \dots, y_T, k}^{(s)}}.$$

These are entries in the unobserved table for iteration $(s + 1)$. They are used as pseudo-data for the M -step of iteration $(s + 1)$. Eventually, the algorithm converges to fitted values for the unobserved table that satisfy mutual independence within each latent class, and such that the corresponding fitted probabilities in the observed table (i.e., added over the latent categories) maximize the log likelihood (14.2). These fitted values also induce ML estimates of the latent class model parameters $\{P(Y_t = y_t | Z = z)\}$ and $\{P(Z = z)\}$.

The EM algorithm is computationally simple and stable. Each iteration increases the likelihood. However, its convergence can be slow. A more problematic issue is that the log likelihood can have local maxima. With either the EM or the Newton–Raphson algorithm, you should perform the fitting process a few times with different starting guesses for the parameter values. The EM algorithm tends to be less sensitive to the choice of starting values. As q increases, multiple local maxima are more likely and the danger also increases of a lack of identifiability.

Standard errors for model parameter estimates result from inverting the model's estimated information matrix. This is a by-product of the Newton–Raphson algorithm but not the EM algorithm. One way to obtain standard errors with the EM algorithm applies a useful formula of Louis (1982) for the observed information. It equals the expected value of the observed information for the loglinear model for the unobserved table minus the expected value of the information for the conditional distribution of Z given the observed data. Lang (1992) gave related results.

Chi-squared statistics comparing observed cell counts to fitted values test the model fit. The residual $df = I^T - q^T(I - 1) - q$, since model (14.1) describes $I^T - 1$ multinomial probabilities using $(I - 1)$ parameters $\{P(Y_t = y_t | Z = z), y_t = 1, \dots, I - 1\}$ at each of q^T combinations of z and t values for the latent variable and the response indicator, and $q - 1$ parameters $\{P(Z = z)\}$. Often, the nature of the variables suggests a value for q , usually quite small (2 to 4). Otherwise, you can start with $q = 2$, and, if the fit is inadequate, increases by steps of 1 as long as the fit shows substantive improvement.

14.1.3 Example: Latent Class Model for Rater Agreement

[Table 14.1](#) shows results for seven pathologists who classified each of 118 slides on the presence or absence of carcinoma in the uterine cervix. For modeling interobserver agreement, the conditional independence assumption of the latent class model is often plausible. With a blind rating scheme, ratings of a given subject or unit by different pathologists are independent. If subjects having true rating in a given category are relatively homogeneous, then ratings by different pathologists may be nearly independent within a given true rating class. Thus, one might posit a latent class model with $q = 2$ classes, one for subjects whose true rating is positive and one for subjects whose true rating is negative. This model expresses the 2^7 joint distribution of the seven ratings as a mixture of two 2^7 distributions, one for each true rating class.

Table 14.1 Diagnoses of Carcinoma and Fits of Latent Class Models^a

Pathologist							Count	Fit		
A	B	C	D	E	F	G		$q = 1$	$q = 2$	$q = 3$
0	0	0	0	0	0	0	34	1.1	23.0	33.8
0	0	0	0	1	0	0	2	1.6	6.6	2.0
0	1	0	0	0	0	0	6	2.2	12.7	6.3
0	1	0	0	0	0	1	1	2.8	1.7	1.5
0	1	0	0	1	0	0	4	3.3	3.6	3.0
0	1	0	0	1	0	1	5	4.2	0.5	4.7
1	0	0	0	0	0	0	2	1.4	3.0	2.1
1	0	1	0	1	0	1	1	1.6	0.2	0.2
1	1	0	0	0	0	0	2	2.8	1.7	1.3
1	1	0	0	0	0	1	1	3.5	0.3	1.6
1	1	0	0	1	0	0	2	4.2	0.5	2.9
1	1	0	0	1	0	1	7	5.3	3.7	6.5
1	1	0	0	1	1	1	1	1.4	2.6	1.4
1	1	0	1	0	0	1	1	1.3	0.1	0.1
1	1	0	1	1	0	1	2	2.0	4.3	2.6
1	1	0	1	1	1	1	3	0.5	3.1	2.0
1	1	1	0	1	0	1	13	3.3	11.5	9.6
1	1	1	0	1	1	1	5	0.9	8.4	8.7
1	1	1	1	1	0	1	10	1.2	13.5	13.6
1	1	1	1	1	1	1	16	0.3	9.9	12.3

^aFits obtained with Latent Gold (Statistical Innovations, Belmont, MA). 1, yes; 0, no.

Source: Based on data in Landis and Koch (1977), not showing empty cells.

[Table 14.2](#) shows results of fitting some latent class models, including another mixture model to be introduced in Section 14.2.5. Because the observed table is sparse, the deviance is mainly useful for comparing models. This is an informal comparison, though, because the chi-squared distribution does not apply for comparing deviances of models with different numbers of latent classes. A model with q classes is a special case of a model with $q^* > q$ classes in which $P(Z = z) = 0$ for $z > q$ and hence falls on the boundary of the parameter space. Ordinary chi-squared likelihood-ratio tests require parameters to fall in the interior of the parameter space [i.e., $0 < P(Z = z) < 1$ for $z = 1, \dots, q^*$]. The actual large-sample null distribution is a mixture of chi-squared distributions (Molenberghs and Verbeke 2007).

Table 14.2 Likelihood-Ratio Statistics for Latent Class Models Fitted to [Table 14.1](#)^a

Number of Latent Classes	Model	Deviance (G^2) Statistic	df
1	Mutual independence	476.8	120
2	Latent class	62.4	112
	Rasch mixture	67.6	118
3	Latent class	15.3	104
	Rasch mixture	27.5	116
4	Latent class	6.4	96
	Rasch mixture (quasi-symmetry)	23.7	114

^aModels fitted with Latent Gold (Statistical Innovations, Belmont, MA).

[Table 14.1](#) also shows the fitted values for latent class models with $q = 1, 2, 3$, for the cells having positive counts. (Each empty cell also has a fitted value, not shown here.) The model with $q = 1$ latent class is the model of mutual independence of the seven ratings. This is equivalent to the loglinear model (Y_1, Y_2, \dots, Y_7) . It fits poorly, as we would expect. With $q = 2$, considerable evidence remains of lack of fit. For instance, the fitted count for a negative rating by each pathologist is 23.0, compared with an observed count of 34. (The small G^2 that [Table 14.2](#) reports for this model does not imply a good fit; from Section 3.2.3, G^2 tends to be highly conservative when most fitted values are very close to 0.) The model with $q = 3$ seems to fit adequately.

Studying the estimated probability $P(Y_t = 1|Z = z)$ of a carcinoma diagnosis for each pathologist, conditional on a given latent class z , helps illuminate the nature of these classes. [Table 14.3](#) reports these for the three-class model. They suggest that (1) the first latent class refers to cases that all pathologists (except occasionally B) agree show no carcinoma; (2) the second latent class refers to cases of strong disagreement, whereby C, D, and F rarely diagnose carcinoma but B, E, and G usually do; and (3) the third latent class refers to cases in which A, B, E, and G agree show carcinoma and C and D usually agree. The estimated proportions in the three latent classes are $\hat{P}(Z = 1) = 0.37$, $\hat{P}(Z = 2) = 0.18$, and $\hat{P}(Z = 3) = 0.45$. The model estimates that 18% of the cases fall in the problematic disagreement class.

A danger with latent variable models, shared by factor analysis for continuous responses, is the temptation to interpret latent variables too literally. For example, here it is tempting to treat latent class 3 as cases truly having carcinoma and a rating of carcinoma given that the subject falls in latent level 3 as being a correct judgment. Realize the tentative nature of the latent variable and be careful not to make the error of reification—treating an abstract construction as if it has actual existence.

Table 14.3 Estimated Probabilities of Diagnosing Carcinoma, for Latent Class Model and Rasch Mixture Model with Three Classes^a

Model	Latent Class	Pathologist						
		A	B	C	D	E	F	G
Latent class	1	0.057	0.138	0.000	0.000	0.055	0.000	0.000
	2	0.513	1.00	0.000	0.058	0.751	0.000	0.631
	3	1.000	0.981	0.858	0.586	1.000	0.476	1.000
Rasch mixture	1	0.022	0.150	0.001	0.000	0.047	0.000	0.022
	2	0.611	0.923	0.052	0.015	0.774	0.009	0.611
	3	0.994	0.999	0.853	0.617	0.997	0.483	0.994

^aResults obtained with Latent Gold (Statistical Innovations, Belmont, MA).

Using the model parameter estimates and Bayes' theorem, we can also estimate $P(Z = z|Y_t = y_t)$ and $P(Z = z|Y_1 = y_1, \dots, Y_T = y_T)$. If a pathologist makes a “yes” rating, for instance, what is the estimated probability that the subject is in the latent class for which agreement on a positive rating usually occurs? We perform further analysis in Section 14.2.6 after studying a simpler model. We could also use methods of Chapter 13, such as a GLMM with a normal rather than categorical latent variable. A logistic-normal random intercept model, for instance, yields subject-specific comparisons of $P(Y_t = 1)$ for various t .

14.1.4 Example: Latent Class Models for Capture–Recapture

We next apply latent class models to capture–recapture modeling for estimating population size. In Section 13.3.4 we used a logistic-normal GLMM for this. With T sampling occasions, a 2^T contingency table displays the data, with scale (captured, not captured) at each occasion. A prediction of the population size equals the prediction for the missing cell count, representing subjects not captured at every occasion, added to the counts in other cells.

With two classes, the latent class model treats the population as a mixture of two types, perhaps determined by genetic or environmental factors. Homogeneity of capture probabilities occurs for subjects within each type, but the type of any given subject is unknown. This model represents a compromise between the mutual independence model, which assumes a single latent class and complete homogeneity, and the logistic-normal GLMM, which assumes a continuous mixture of capture probabilities rather than two classes.

We illustrate with the data set on snowshoe hares in [Table 13.6](#), having $T = 6$ captures. The model of mutual independence predicts that $\hat{N} = 75$. Its 95% profile likelihood confidence interval for N is $(70, 83)$. The latent class model with two classes has $\hat{N} = 85$ and a profile likelihood interval of $(74, 106)$. The latent class model with three classes gives similar results. The logistic-normal GLMM in Section 13.3.4 gave the interval $(75, 154)$, so these seem too short to be trusted. This simple latent class model may not capture all the existing heterogeneity.

Possible models other than latent class or parametric random effects models include loglinear models (Cormack 1989). They are marginal models, applying to probabilities averaged over subjects. Let Y_t denote the binary capture variable for a randomly selected subject at occasion t . The simplest model, (Y_1, Y_2, \dots, Y_T) , treats capture events as mutually independent and is equivalent to the logistic-normal model [\(13.12\)](#) with $\sigma = 0$ and latent class model [\(14.1\)](#) with $q = 1$. The loglinear model $(Y_1 Y_2, Y_1 Y_3, \dots, Y_{T-1} Y_T)$ allows an association between pairs of capture variables. Alternatively, a simpler model with Markov structure $(Y_1 Y_2, Y_2 Y_3, \dots, Y_{T-1} Y_T)$ or with the same association for each pair of occasions may be useful (Exercise 14.3).

In capture–recapture experiments, \hat{N} and confidence intervals for N depend strongly on the choice of model. Standard goodness-of-fit criteria are of limited help. Two models can fit the observed counts well, yet yield quite different predictions for the unobserved count. For instance, for the snowshoe hare data, the loglinear models of mutual independence and of two-factor association both fit relatively well ($G^2 = 58.3$, $df = 56$ for mutual independence and $G^2 = 32.4$, $df = 41$ for the two-factor model); however, their \hat{N} values are 75 and 105.

Simpler models usually give narrower confidence intervals for N , through the usual benefits of model parsimony. This is not necessarily good for this type of application. A narrow confidence interval for N is desirable, but not at the expense of severe sacrifice in the actual confidence level. Intervals based on a possibly unrealistic assumption of subject homogeneity are often overly optimistic. Simulations suggest that actual coverage probabilities can then be well below nominal levels when even slight model misspecification occurs. Allowance for heterogeneity among subjects results in wider intervals. Severe population heterogeneity makes reaching useful conclusions difficult, as intervals can be very wide (Burnham and Overton 1978, Coull and Agresti 1999).

14.1.5 Example: Latent Class Transitional Models

The basic latent class model has been generalized in many ways. For example, Reboussin and Ialongo (2010) modeled drug use among high school students who suffer from attention deficit hyperactivity disorder (ADHD). Their model consists of two separate latent class models: A longitudinal latent transition model has latent classes that are stages of marijuana use and describes the probability of transitioning between the stages. A cross-sectional latent class predictor model empirically constructs ADHD subtypes and describes the influence of those subtypes on the transition rates.

Other generalizations that focus on transitions use continuous latent variables and resemble multivariate random effects models. For example, in a longitudinal aging study, Lin et al. (2008) modeled repeated transitions between independence and disability states of activities of daily living. Their multistate transition model is designed for the analysis of repeated episodes of multiple states representing different health status, where some states (such as death) are absorbing. Transitions among multiple states are modeled jointly using multivariate latent variables. A state-specific latent variable represents an individual's tendency to remain in a nonabsorbing state, beyond the time explained by covariates, and to account for correlation among repeated sojourns in the same state. Correlation among sojourns across different states is accounted for by the correlation between the different latent variables.

14.2 NONPARAMETRIC RANDOM EFFECTS MODELS

In spite of its popularity and attractive features, the normality assumption for random effects in ordinary GLMMs can rarely be closely checked. McCulloch and Neuhaus (2011) noted that distributions of predicted values are highly dependent on their assumed distribution and are not reliable indicators of the true random effects distribution. An obvious concern of this or any parametric assumption for the random effects is possibly harmful effects of misspecification. To check sensitivity to this assumption, we can fit GLMMs using alternative or more general random effects assumptions.

14.2.1 Logistic Models with Unspecified Random Effects Distribution

A nonparametric approach (Aitkin 1999, Heckman and Singer 1984) guards against possibly harmful misspecification effects. This uses an unspecified random effects distribution on a finite set of mass points. The location of the mass points and their probabilities are parameters. The number of mass points can be fixed. When this number is itself unknown, we treat it as fixed in the estimation process but increase it sequentially until the likelihood is maximized. The maximization usually requires relatively few mass points. Even allowing a continuous mixture distribution, the nonparametric estimate of that distribution takes a finite number of points (e.g., Lindsay et al. 1991). In fact, fitting a model having only two mass points often results in fixed effects estimates quite similar to those with the full maximization. This approach is useful primarily when the random effects distribution is not itself of direct interest, since the nonparametric estimate of that distribution tends to be poor even for very large samples.

Model fitting is actually simpler than for models with normal random effects, since the integral that determines the likelihood function simplifies to a finite sum. However, this approach also has disadvantages. For instance, with multivariate random effects it cannot provide simple correlation structure as the normal can. Also, the ML estimate of the random effects distribution often places some weight at $\pm\infty$. Although this can be useful with binary data for identifying a subsample for which the estimated response probability equals 1 or equals 0 for all observations in a cluster, it is not then possible to describe heterogeneity with an estimated variance component.

14.2.2 Example: Attitudes About Legalized Abortion

To illustrate this approach, we reanalyze [Table 13.3](#) on attitudes about legalized abortion. In Section 13.3.2 we fitted the logistic-normal model,

$$(14.3) \text{ logit}[P(Y_{it} = 1|u_i)] = \alpha + \beta_i + \gamma x_i + u_i,$$

with x_i = gender (1 = female) and parameters $\{\beta_i\}$ representing three conditions under which abortion might be legal.

Treating u_i instead nonparametrically, the likelihood maximizes with a two-point mixture distribution. Estimated abortion item effects are $\hat{\beta}_1 - \hat{\beta}_3 = 0.83$ ($SE = 0.16$), $\hat{\beta}_2 - \hat{\beta}_3 = 0.30$ ($SE = 0.16$), and $\hat{\beta}_1 - \hat{\beta}_2 = 0.52$ ($SE = 0.16$). Results are similar to those in [Table 13.3](#) for the normal random effects approach.

14.2.3 Example: Nonparametric Mixing of Logistic Regressions

Follman and Lambert (1989) analyzed the effect of the dosage of a poison on the probability of death of a protozoan of a particular genus. [Table 14.4](#) shows the data. They assumed two unobserved types of that genus.

[Table 14.4](#) Number of Protozoa Exposed to Poison Dose and Number That Died

Poison Dose	Exposed	Dead	Poison Dose	Exposed	Dead
4.7	55	0	5.1	53	22
4.8	49	8	5.2	53	37
4.9	60	18	5.3	51	47
5.0	55	18	5.4	50	50

Source: Follman and Lambert (1989). Reprinted with permission from the *Journal of the American Statistical Association*.

Let $\pi_i(x)$ denote the probability of death at log dose level x for genus type i , $i = 1, 2$. Let ρ denote the probability a protozoan belongs to genus type 1. Their model specifies

$$\pi(x) = \rho\pi_1(x) + (1 - \rho)\pi_2(x), \quad \text{where } \text{logit}[\pi_i(x)] = \alpha_i + \beta x,$$

with unknown ρ . The curve for $\pi(x)$ is a weighted average of two curves having the same logistic shapes but different intercepts.

The ordinary logistic regression model is the special case $\rho = 1$. Its fit, $\text{logit}[\hat{\pi}(x)] = -68.4 + 42.1x$, is poor, with deviance $G^2 = 24.7$ ($\text{df} = 6$). The fit of the mixture model is

$$\hat{\pi}(x) = 0.34\hat{\pi}_1(x) + 0.66\hat{\pi}_2(x), \quad \text{with}$$

$$\text{logit}[\hat{\pi}_1(x)] = -196.2 + 124.8x, \quad \text{logit}[\hat{\pi}_2(x)] = -205.7 + 124.8x.$$

[Figure 14.2](#) shows the fit. This is much better, with $G^2 = 3.4$ ($\text{df} = 4$); that is, double the maximized log-likelihood increases by 21.3 by adding two parameters: an additional intercept and the probability for the mixture. Follman and Lambert noted that with eight dose levels, at most two mixture points are identifiable for this model.

[Figure 14.2](#) Fit of binary mixture of logistic regressions to [Table 14.4](#) [model fitted using Latent Gold (Statistical Innovations, Belmont, MA)].

Pathologist	F	D	C	A	G	E	B
Estimate	-3.70	-3.15	-1.87	1.48	1.48	2.26	3.52
Comparison	—	—	—	—	—	—	—

The ordinary GLMM assumes a normal mixture of logistic curves. It gives a deviance reduction of only 1.7 compared to the ordinary logistic model with $\rho = 1$.

14.2.4 Is Misspecification of Random Effects a Serious Problem?

Is it worth the trouble to consider alternatives to the normality assumption for random effects in GLMMs, whether they be parametric or nonparametric? For logistic random intercept models, different assumptions for the random effects distribution often provide similar results for estimating the regression effects. Choosing an incorrect random effects distribution does not tend to bias estimators of those effects. The true distribution for the random effects being skewed can result in some bias for the normal intercept estimator (Neuhaus et al. 1992). The choice of random effects distribution also usually has little impact on efficiency of estimation. Also, using a nonparametric approach when the true distribution is normal does not result in much efficiency loss (Neuhaus and Lesperance 1996).

When the true random effects distribution is far from normal, there can be some efficiency loss for the logistic-normal estimator. One such case is when the true distribution is a two-point mixture with large variance component, such as suggested in the previous example. Agresti et al. (2004) studied this with various models, such as a simple one-way random effects model. In cluster i , let y_{it} be a Bernoulli variate satisfying

$$(14.4) \quad \text{logit}[P(Y_{it} = 1|u_i)] = \alpha + u_i, \quad i = 1, \dots, n, \quad t = 1, \dots, T,$$

where $\text{var}(u_i) = \sigma^2$. Simulated samples from this model used various n , T , α , and σ , and various true distributions for u_i including normal, uniform, exponential, and binary. When the true distribution is a two-point mixture, the normal approach loses efficiency in estimating $\{\mu_i = P(Y_{it} = 1|u_i)\}$, more so as σ and T increase. For example, when $n = T = 30$, $\alpha = 0$, and the mixture has probability 0.50 at each point, the expected value of $|\hat{\mu}_i - \mu_i|$ is (0.061, 0.023) for the (normal, nonparametric) approach when $\sigma = 1.0$, and (0.045, 0.013) when $\sigma = 2.0$.

The example from Follman and Lambert (1989) discussed in Section 14.2.3, which has a covariate but $T = 1$, illustrates the potential efficiency loss with the logistic-normal GLMM. The two-point mixture model has $\hat{\beta} = 124.8$ with $SE = 25.2$, for which $\hat{\beta}/SE = 4.9$. The normal mixture model has $\hat{\beta} = 65.5$ with $SE = 19.5$, for which $\hat{\beta}/SE = 3.4$.

Some research suggests that the random effects distribution has to be highly nonnormal for the normal GLMM to suffer in bias or efficiency. McCulloch and Neuhaus (2011) noted that the accuracy of predicted random effects is not much affected by mild-to-severe violations of the assumed structure. Assuming different distributions for the random effects can yield quite different predicted values yet have similar performance in terms of overall accuracy of prediction; in fact, they noted that a significantly better fitting random effects distribution may not perform better for prediction. However, Heagerty and Zeger (2000) noted that other types of misspecification can be more crucial. Regarding bias, they argued that sensitivity to the random effects assumption is greater for estimating regression parameters in random effects models than estimating their counterparts in corresponding marginal models. They illustrated this with a model violation by which the variance of the random effects depends on values of covariates. They concluded that between-cluster effects may be more sensitive to correct specification of the random effects distribution than within-cluster effects. This is an advantage of using marginal models for between-cluster effects.

14.2.5 Rasch Mixture Model

From Section 13.1.4, for subject i with item t the Rasch model for a binary response is

$$(14.5) \quad \text{logit}[P(Y_{it} = 1|u_i)] = \alpha + \beta_t + u_i, \quad t = 1, \dots, T$$

with a constraint on $\{\beta_t\}$. The GLMM treats $\{u_i\}$ as normal random effects. Lindsay et al. (1991) studied this model when u_i instead can assume only a finite number q of values. Denote its distribution by

$$P(U = a_k) = \rho_k, \quad k = 1, \dots, q,$$

for unknown $\{a_k\}$ and $\{\rho_k\}$ satisfying a constraint for identifiability, such as $\sum_k \rho_k a_k = 0$. This model is called a *Rasch mixture model*. As in other random effects models, u_i is unobserved, and the T responses are assumed conditionally independent at each fixed u_i value. It differs from the ordinary latent class model for binary responses having q latent classes (Section 14.1), since it assumes structure (14.5) for $P(Y_{it} = 1|u_i)$ whereas latent class model (14.1) assumes no structure for $P(Y_t = y_t|Z = z)$.

For the Rasch mixture model, the marginal probability of a sequence of responses (y_1, \dots, y_T) is

$$\pi_{y_1, \dots, y_T} = \sum_{k=1}^q \rho_k \left[\prod_{t=1}^T \frac{\exp[y_t(\beta_t + a_k)]}{1 + \exp(\beta_t + a_k)} \right].$$

Substituting this in the multinomial log likelihood (14.2), we can estimate $\{a_k, \rho_k\}$ and $\{\beta_t\}$ using Newton–Raphson or EM algorithms. As q increases, the maximized likelihood increases and the fit improves. However, Lindsay et al. (1991) showed that, with T items, the likelihood no longer changes once $q = (T + 1)/2$. Then, the model gives the same fit to the 2^T observed table as the quasi-symmetry model (11.32). Thus, this simpler latent class model has a symmetric conditional association structure among the observed variables.

14.2.6 Example: Modeling Rater Agreement Revisited

For the ratings of carcinoma by seven pathologists (Table 14.1), Table 14.2 also summarizes the fit of Rasch mixture models. Here, $P(Y_{it} = 1|u_i)$ in (14.5) denotes the probability of a carcinoma diagnosis for pathologist t evaluating slide i . With $q = 3$, it does not fit significantly more poorly than the latent class model. With $T = 7$ raters, the discrete mixture can take at most $(T + 1)/2 = 4$ points. The model with $q = 4$ is equivalently the quasi-symmetry model. It does not seem to fit better than with $q = 3$.

Figure 14.3 shows $\{\hat{\beta}_t\}$ for the Rasch mixture model with $q = 3$, setting $\sum_t \hat{\beta}_t = 0$. These describe variation among the pathologists' response distributions at each latent level. For a given latent class, for instance, the estimated odds of a carcinoma diagnosis for pathologist B are $\exp(3.52 - 1.48) = 7.7$ times the estimated odds for pathologist A. Pathologist B tends to make a carcinoma diagnosis most often, and D and F the least. The figure also shows results of a 90% Bonferroni comparison of the 21 pairs of pathologists, based on Wald intervals for all pairwise differences $\hat{\beta}_t - \hat{\beta}_s$.

Figure 14.3 Pathologist estimates for Rasch mixture model and results of 90% Bonferroni simultaneous comparison.

Pathologist	F	D	C	A	G	E	B
Estimate	-3.70	-3.15	-1.87	1.48	1.48	2.26	3.52
Comparison	—	—	—	—	—	—	—

For pathologist t , conditional on latent level k for a slide,

$$\exp(\hat{\beta}_t + \hat{a}_k)/[1 + \exp(\hat{\beta}_t + \hat{a}_k)]$$

estimates the probability of a carcinoma diagnosis. Table 14.3 reports these, which use $\hat{a}_1 = -5.25$, $\hat{a}_2 = -1.02$, and $\hat{a}_3 = 3.63$. They are similar to the estimates for the ordinary latent class model with $q = 3$ but a bit smoother, with fewer estimates at the boundary. Again, at latent level 1 pathologists tend not to diagnose carcinoma, at level 2 many disagreements occur, and at level 3 pathologists tend to diagnose carcinoma. The estimated latent class proportions are $\hat{p}_1 = 0.37$, $\hat{p}_2 = 0.19$, and $\hat{p}_3 = 0.43$, similar to the ordinary latent class model.

Model (14.5) implies that the association between each Y_t and U has log odds ratio $(a_k - a_l)$ for levels k and l of U . For instance, in the third latent class the estimated odds that a pathologist diagnoses carcinoma are $\exp[3.63 - (-5.25)] > 7000$ times those in the first latent class. The large $\{\hat{a}_k - \hat{a}_l\}$ suggest strong association between each pathologist's rating and the latent variable. This induces strong association between pairs of pathologist ratings. The model-fitted odds ratios between pairs of raters vary between about 7 and 400, but confidence intervals reveal that these estimates are very imprecise. However, the quite varied $\{\hat{\beta}_t\}$ suggest that substantial marginal heterogeneity exists among the seven ratings. This causes heterogeneity in pairwise levels of agreement.

The mutual independence model is the special case of the Rasch mixture model with $q = 1$; that is, $\rho_1 = 1$. For Table 14.1 the Rasch mixture model with $q = 3$ has only four more parameters than the mutual independence model (i.e., ρ_k and a_k , $k = 1, 2$). Yet it fits well and has simple interpretations.

14.2.7 Nonparametric Mixtures and Quasi-symmetry

A distribution-free approach for u_i with the Rasch form of model (14.5) implies the quasi-symmetry loglinear model marginally (Darroch 1981, Tjur 1982). Let Y_i denote the sequence of T responses for subject i . For the possible outcomes $y = (y_1, \dots, y_T)$, where each $y_t = 1$ or 0, and removing α and the constraint on $\{\beta_t\}$,

$$\begin{aligned} P(Y_i = y|u_i) &= \prod_t \left[\frac{\exp(\beta_t + u_i)}{1 + \exp(\beta_t + u_i)} \right]^{y_t} \left[\frac{1}{1 + \exp(\beta_t + u_i)} \right]^{1-y_t} \\ &= \frac{\exp[u_i (\sum_t y_t) + \sum_t y_t \beta_t]}{\prod_t [1 + \exp(\beta_t + u_i)]}. \end{aligned}$$

Let F denote the cdf of u_i . The marginal probability of sequence y for a randomly selected subject is (suppressing the subject label)

$$\pi_{y_1, \dots, y_T} = E_U P(Y = y|U) = \exp \left(\sum_t y_t \beta_t \right) \int \frac{\exp[u (\sum_t y_t)]}{\prod_t [1 + \exp(\beta_t + u)]} dF(u).$$

This probability contributes to the log likelihood, which is (14.2) for a multinomial distribution over the 2^T cells for possible y . Regardless of the choice for F , the integral is complex. However, it depends on the data only through $\sum_t y_t$. A more general model replaces this integral by a separate parameter for each value of $\sum_t y_t$. This model has form

$$(14.6) \quad \log \pi_{y_1, \dots, y_T} = \sum_t y_t \beta_t + \lambda_{y_1+...+y_T}.$$

The final term represents a separate parameter at each value of $\sum_t y_t$.

The implied marginal model (14.6) has interaction term that is invariant to any permutation of the response outcomes y , since each such permutation yields the same sum, $\sum_t y_t$. Thus, it is the loglinear model of quasi-symmetry (11.32). No matter what form F takes, the marginal model has the same main-effect structure, and it has an interaction term that is a special case of the one in (14.6). Thus, we can consistently estimate $\{\beta_t\}$ using the ordinary ML estimates for the quasi-symmetry model. In fact, Tjur (1982) showed that these estimates are also the conditional ML estimates, treating $\{u_i\}$ as fixed effects and conditioning on their sufficient statistics. The interaction parameters in model (14.6) result from the dependence in responses among variables, due to heterogeneity in $\{u_i\}$.

14.2.8 Example: Attitudes About Legalized Abortion Revisited

We illustrate for the opinions about legalized abortion analyzed with a GLMM in Section 13.3.2 and with a nonparametric random effects approach in Section 14.2.2. For model (14.3), estimated within-subject comparisons $\beta_t - \beta_s$ of items result from fitting a quasi-symmetric loglinear model. Let $\mu_g(y_1, y_2, y_3)$ denote the expected frequency for gender g making response y_t to item t , $t = 1, 2, 3$, where for item t , $y_t = 1$ for approval and 0 for disapproval. The loglinear model is

$$(14.7) \quad \log \mu_g(y_1, y_2, y_3) = \beta_1 y_1 + \beta_2 y_2 + \beta_3 y_3 + \beta_4 g + \lambda_{y_1+y_2+y_3}.$$

For $y_1 + y_2 + y_3 = k$, λ_k refers to all cells in which subjects voiced approval for k of the three items, $k = 0, 1, 2, 3$. The ML fit, which has $G^2 = 10.2$ with $df = 9$, yields $\hat{\beta}_1 - \hat{\beta}_2 = 0.521$ ($SE = 0.154$), $\hat{\beta}_1 - \hat{\beta}_3 = 0.828$ ($SE = 0.160$), and $\hat{\beta}_2 - \hat{\beta}_3 = 0.307$ ($SE = 0.161$). These are similar to the GLMM estimates (Table 13.3) and nonparametric random effects model estimates in Section 14.2.1. They also are the conditional ML estimates for model (14.3), treating $\{u_i\}$ as fixed. With this approach or conditional ML, however, we cannot estimate between-groups effects, such as the gender effect in model (14.3). [The β_4 parameter in model (14.7) refers to relative sample sizes of males and females and is not the same as the γ gender effect in (14.3).]

14.3 BETA-BINOMIAL MODELS

The beta-binomial model is a parametric mixture model that is another alternative to binary GLMMs with normal random effects. As with other mixture models that assume a binomial distribution at a fixed parameter value, the marginal distribution permits more variation than the binomial. Thus, a model using the beta-binomial can handle overdispersion occurring with ordinary binomial models.

14.3.1 Beta-Binomial Distribution

The beta-binomial distribution results from a beta distribution mixture of binomials. Suppose that (a) given π , Y has a binomial distribution, $\text{bin}(n, \pi)$, and (b) π has a beta distribution. The beta pdf (Sec. 1.6.2) is

$$(14.8) \quad f(\pi; \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \pi^{\alpha_1-1} (1-\pi)^{\alpha_2-1}, \quad 0 \leq \pi \leq 1,$$

with parameters $\alpha_1 > 0$ and $\alpha_2 > 0$, for the gamma function $\Gamma(\cdot)$. Let

$$\mu = \frac{\alpha_1}{\alpha_1 + \alpha_2}, \quad \theta = 1/(\alpha_1 + \alpha_2).$$

The beta distribution for π has mean and variance

$$E(\pi) = \mu, \quad \text{var}(\pi) = \mu(1-\mu)\theta/(1+\theta).$$

Marginally, averaging with respect to the beta distribution for π , Y has the *beta-binomial distribution*. Its mass function is

$$p(y; \alpha_1, \alpha_2) = \binom{n}{y} \frac{B(\alpha_1 + y, n + \alpha_2 - y)}{B(\alpha_1, \alpha_2)}, \quad y = 0, 1, \dots, n,$$

for the beta function $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$. In terms of μ and θ , the beta-binomial mass function is

$$(14.9) \quad p(y; \mu, \theta) = \binom{n}{y} \frac{[\prod_{k=0}^{y-1} (\mu + k\theta)][\prod_{k=0}^{n-y-1} (1-\mu + k\theta)]}{\prod_{k=0}^{n-1} (1+k\theta)}.$$

It is easier to understand the nature of this distribution from its moments than from its mass function. The first two moments are

$$E(Y) = n\mu, \quad \text{var}(Y) = n\mu(1-\mu)[1+(n-1)\theta/(1+\theta)].$$

In fact, $\theta/(1+\theta) = 1/(\alpha_1 + \alpha_2 + 1)$ is the correlation between each pair of the individual Bernoulli random variables that sum to Y .

As $\theta \rightarrow 0$ in the beta distribution, $\text{var}(\pi) \rightarrow 0$ and that distribution converges to a degenerate distribution at μ . Then $\text{var}(Y) \rightarrow n\mu(1-\mu)$ and the beta-binomial distribution converges to the $\text{bin}(n, \mu)$.

14.3.2 Models Using the Beta-Binomial Distribution

Models using the beta-binomial distribution permit μ , and hence $E(Y)$, to depend on explanatory variables. The simplest models let θ be the same unknown constant for all observations. More general models let θ depend on covariates, such as by allowing a different θ for each group of interest (Prentice 1986). Models can use any of the usual link functions for binary data, but the logit is most common. For observation i with n_i trials, assuming that y_i has a beta-binomial distribution with index n_i and parameters (μ_i, θ) , the model links μ_i to explanatory variables by

$$\text{logit}(\mu_i) = \alpha + \boldsymbol{\beta}^T \mathbf{x}_i.$$

The beta-binomial is not in the natural exponential family, even for known θ . Articles using beta-binomial models have employed a variety of fitting methods (Note 14.4), including Newton–Raphson.

14.3.3 Quasi-likelihood with Beta-Binomial Type Variance

A related but simpler approach for overdispersed binary counts uses quasi-likelihood with similar variance function as the beta-binomial. The quasi-likelihood variance function is

$$(14.10) \quad v(\mu_i) = n_i \mu_i (1 - \mu_i) [1 + (n_i - 1)\rho]$$

with $|\rho| \leq 1$. Although motivated by the beta-binomial model with its correlation between binary components, this variance function results merely from assuming that π_i has a distribution with $\text{var}(\pi_i) = \rho \mu_i (1 - \mu_i)$.

This variance function also results from assuming a common correlation ρ between each pair of the n_i individual Bernoulli random variables that sum to y_i , without specifically assuming a beta mixture (Altham 1978). Suppose that $\pi_i = P(Y_{it} = 1) = 1 - P(Y_{it} = 0)$, for $t = 1, \dots, n_i$, and $\text{corr}(Y_{is}, Y_{it}) = \rho$ for $s \neq t$. Then, $\text{var}(Y_{it}) = \pi_i(1 - \pi_i)$, $\text{cov}(Y_{is}, Y_{it}) = \rho \pi_i(1 - \pi_i)$, and

$$\begin{aligned} \text{var}\left(\sum_t Y_{it}\right) &= \sum_t \text{var}(Y_{it}) + 2 \sum_{s < t} \text{cov}(Y_{is}, Y_{it}) \\ &= n_i \pi_i (1 - \pi_i) + n_i(n_i - 1) \rho \pi_i (1 - \pi_i) = n_i \pi_i (1 - \pi_i) [1 + \rho(n_i - 1)]. \end{aligned}$$

The ordinary binomial variance results when $\rho = 0$. Overdispersion occurs when $\rho > 0$.

For this quasi-likelihood approach, Williams (1982) proposed an iterative routine for estimating β and the overdispersion parameter ρ . He let $\hat{\rho}$ be such that the resulting Pearson X^2 that sums the squared Pearson residuals for this variance function equals the residual df for the model. This requires an iterative two-step process of (1) solving the quasi-likelihood equations for β for a given $\hat{\rho}$, and then (2) using the updated $\hat{\beta}$ solving for $\hat{\rho}$ in the equation that equates X^2 (which depends on $\hat{\beta}$ and $\hat{\rho}$) to its df.

An alternative quasi-likelihood approach, presented in Section 4.7.3, uses the simpler inflated binomial variance function

$$(14.11) \quad v(\mu_i) = \phi n_i \mu_i (1 - \mu_i).$$

The ordinary binomial variance has $\phi = 1.0$ and overdispersion occurs when $\phi > 1$. With this approach, $\hat{\beta}$ is the same as its ML estimate for the ordinary binomial model. Commonly, $\hat{\phi} = X^2/\text{df}$, where X^2 is the Pearson fit statistic for the binomial model (Finney 1947). The standard errors for the overdispersion approach multiply those for the binomial model by $\hat{\phi}^{1/2}$.

Liang and McCullagh (1993) showed several examples using these two variance functions. A plot of the standardized residuals for the ordinary binomial model against the indices $\{n_i\}$ can provide insight about which is more appropriate. When the residuals show an increasing trend in their spread as n_i increases, the beta-binomial-type variance function may be more appropriate. This is because when the beta-binomial variance holds, the residuals from an ordinary binomial model have denominator that is progressively too small as n_i increases. The two quasi-likelihood approaches are equivalent when $\{n_i\}$ are identical. Only when the indices vary considerably might results differ much. Because the variance function $v(\mu_i) = \phi n_i \mu_i (1 - \mu_i)$ has a structural problem when $n_i = 1$ (Section 4.7.3) and has less direct motivation, we prefer quasi-likelihood with the beta-binomial variance function.

14.3.4 Example: Teratology Overdispersion Revisited

[Table 4.7](#) showed results of a teratology experiment. Female rats on iron-deficient diets were assigned to four groups. Group 1 was given only placebo injections. The other groups were given injections of an iron supplement according to various schedules. The rats were made pregnant and then sacrificed after 3 weeks. For each fetus in each rat's litter, the response was whether the fetus was dead. Because of unmeasured covariates, it is natural to permit the probability of death to vary from litter to litter within a particular treatment group.

Let y_i denote the number dead out of the n_i fetuses in litter i . Let π_{it} denote the probability of death for fetus t in litter i . We use the model

$$\text{logit}(\pi_{it}) = \alpha + \beta_2 z_{2i} + \beta_3 z_{3i} + \beta_4 z_{4i},$$

where $z_{gi} = 1$ if litter i is in group g and 0 otherwise.

First, suppose that y_i is a $\text{bin}(n_i, \pi_{it})$ variate, independent from litter to litter. This treats all litters in a group g as having the same probability of death, $\exp(\alpha + \beta_g) / [1 + \exp(\alpha + \beta_g)]$, where $\beta_1 = 0$. The ML estimates are $\hat{\alpha} = 1.14$ ($SE = 0.13$), $\hat{\beta}_2 = -3.32$ ($SE = 0.33$), $\hat{\beta}_3 = -4.48$ ($SE = 0.73$), $\hat{\beta}_4 = -4.13$ ($SE = 0.48$). However, this binomial ML approach has evidence of overdispersion, with $X^2 = 154.7$ and $G^2 = 173.5$ ($df = 54$).

By contrast, [Table 14.5](#) shows ML estimates and standard errors for the beta-binomial model, which permits heterogeneity for litters in a group. The overdispersion results in inflated SE values compared with binomial ML. For the beta-binomial fit, $\theta/(1 + \theta) = 0.241$, so the fit treats the variance of Y_i as

Table 14.5 Estimates for Several Logistic Models Fitted to [Table 4.7](#)

Parameter	Type of Logistic Model ^a				
	Beta-bin. ML	QL(1)	QL(2)	GEE	GLMM
Intercept	1.35 (0.24)	1.21 (0.22)	1.21 (0.27)	1.14 (0.28)	1.80 (0.36)
Group 2	-3.11 (0.52)	-3.37 (0.56)	-3.32 (0.56)	-3.37 (0.43)	-4.51 (0.74)
Group 3	-3.87 (0.86)	-4.59 (1.30)	-4.48 (1.24)	-4.58 (0.62)	-5.86 (1.19)
Group 4	-3.92 (0.68)	-4.25 (0.85)	-4.13 (0.81)	-4.25 (0.60)	-5.59 (0.92)
Overdispersion	$\frac{\hat{\theta}}{1 + \hat{\theta}} = 0.241$	$\hat{\rho} = 0.192$	$\hat{\phi} = 2.86$	$\hat{\rho} = 0.185$	$\hat{\sigma} = 1.53$

^aQL is quasi-likelihood with (1) beta-binomial-type variance, (2) inflated binomial variance; GEE uses exchangeable working correlations. Values in parentheses are standard errors.

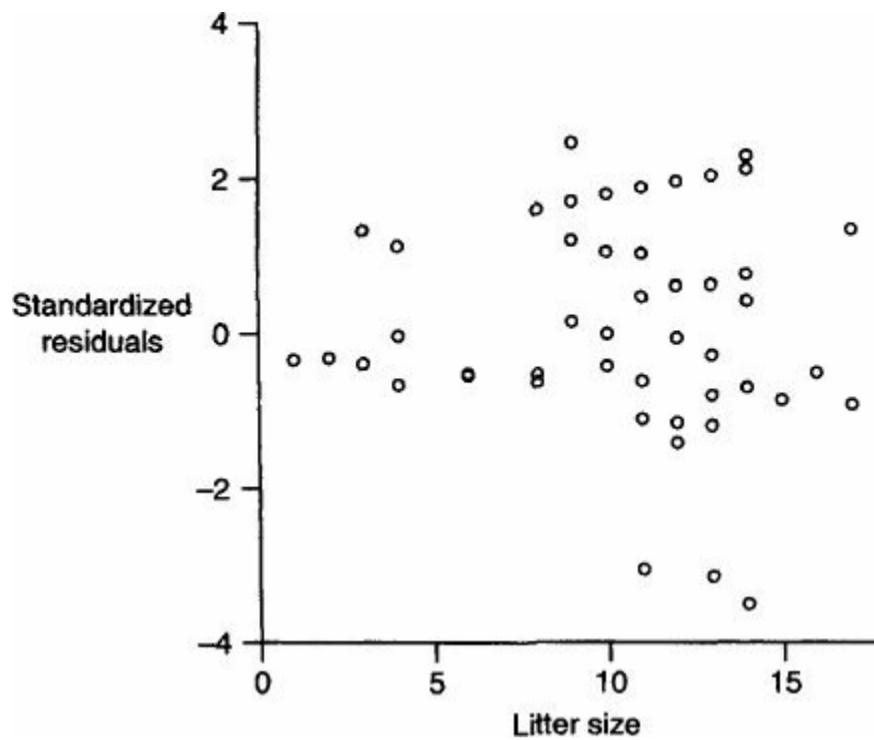
$$\text{var}(Y_i) = n_i \mu_i (1 - \mu_i) [1 + 0.241(n_i - 1)].$$

This corresponds roughly to a doubling of the variance relative to the binomial with a litter size of 5 and a tripling with $n_i = 9$.

[Table 14.5](#) also shows results for the two quasi-likelihood approaches. Estimates and standard errors are qualitatively similar. For variance function $v(\mu_i) = \phi n_i \mu_i (1 - \mu_i)$, the estimates equal the binomial ML estimates but SE values are multiplied by $\hat{\phi}^{1/2} = \sqrt{X^2/\text{df}} = \sqrt{154.7/54} = 1.69$.

[Figure 14.4](#) plots the standardized residuals against litter size for the binomial logit model. The apparent increase in their variability as litter size increases suggests that the beta-binomial variance function is plausible. For that variance function, the probabilities of death for litters of a particular group have standard deviation $\sqrt{\rho \mu_i (1 - \mu_i)}$, where in the beta-binomial distribution ρ corresponds to $\theta/(1 + \theta)$. For the QL fit, $\hat{\rho} = 0.192$, so this standard deviation equals 0.22 when the mean is 0.50 and 0.13 when the mean is 0.10 or 0.90. This is considerable heterogeneity. More generally, a model could let ρ vary by treatment group or be different for the placebo group from the others. We leave this to the reader.

Figure 14.4 Standardized Pearson residuals for binomial logistic model fitted to [Table 4.7](#).



For comparison, [Table 14.5](#) also shows results with the GEE approach to fitting the logistic model, assuming exchangeable working correlation structure for observations within a litter. The empirical sandwich adjustment increases the SE values compared with binomial ML. The estimated within-litter correlation between the binary responses is 0.185. This is comparable to the value of 0.192 that yields the quasi-likelihood results with beta-binomial variance function. The GEE standard errors are somewhat different from those with the quasi-likelihood approach. It may be that the sample size is insufficient for the GEE sandwich adjustment, which tends to underestimate standard errors unless the number of clusters is quite large. Or, this may merely reflect the different variance function for the GEE approach.

Finally, [Table 14.5](#) shows results for the GLMM that adds a normal random intercept u_i for litter i to the binomial logistic model. Estimated effects are larger for this logistic-normal model, since they are cluster-specific (for cluster = litter) rather than population-averaged. Even with all these adjustments for overdispersion, [Table 14.5](#) shows that strong evidence remains that the probability of death is substantially lower for each treatment group than the placebo group.

14.3.5 Conjugate Mixture Models

The beta-binomial model is an example of a *conjugate mixture model*. These are models for which the marginal distribution has closed form. The data have a particular distribution, conditional on a parameter, and then the parameter has its own distribution such that the marginal distribution has closed form.

Likewise, from Section 1.6.2, in Bayesian methods the conjugate prior distribution is a distribution that, when combined with the likelihood, gives a closed form for the posterior distribution. For instance, for binomial observations with beta prior distribution for the parameter, the posterior distribution is also beta.

Next, we present a conjugate mixture model for count data. It uses a gamma distribution to mix the Poisson parameter. A disadvantage of the conjugate mixture approach is the lack of generality and flexibility, requiring a different mixture distribution for each type of problem. In addition, the extra variability need not enter on the same scale as the ordinary predictors, and it can be difficult to have multivariate random effects structure. Lee et al. (2006) discussed the conjugate approach and discussed a variety of hierarchical models of GLMM form in which the random effects need not be normal.

14.4 NEGATIVE BINOMIAL REGRESSION

The *negative binomial* is a conjugate mixture distribution for count data. It is useful when overdispersion occurs with Poisson GLMs.

14.4.1 Gamma Mixture of Poissons Is Negative Binomial

A severe limitation of Poisson models is that the variance of Y must equal the mean (Section 4.3.3). Hence, at a fixed mean the variance cannot decrease as additional predictors enter the model. Count data often show overdispersion, with the variance exceeding the mean. This might happen, for instance, because some relevant explanatory variables are not in the model. A mixture model is a flexible way to account for overdispersion. At a fixed setting of the predictors used, given the mean the distribution of Y is Poisson, but the mean itself varies according to some distribution.

Suppose that (1) given λ , Y has a Poisson distribution with mean λ , and (2) λ has a gamma distribution, $G(k, \mu)$. The gamma probability density function for λ is

$$(14.12) \quad f(\lambda; k, \mu) = \frac{(k/\mu)^k}{\Gamma(k)} \exp(-k\lambda/\mu) \lambda^{k-1}, \quad \lambda \geq 0.$$

This gamma distribution has

$$E(\lambda) = \mu, \quad \text{var}(\lambda) = \mu^2/k.$$

The parameter $k > 0$ describes the shape. The density is skewed to the right, but the degree of skewness (which equals $2/\sqrt{k}$) decreases as k increases.

Marginally, the gamma mixture of the Poisson distributions yields the negative binomial distribution for Y (Greenwood and Yule 1920). Its probability mass function is

$$(14.13) \quad p(y; k, \mu) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{k}{\mu+k}\right)^k \left(1 - \frac{k}{\mu+k}\right)^y, \quad y = 0, 1, 2, \dots$$

In terms of the dispersion parameter $\lambda = 1/k$,

$$E(Y) = \mu, \quad \text{var}(Y) = \mu + \gamma\mu^2.$$

The greater γ , the greater the overdispersion relative to the Poisson. As $\gamma \rightarrow 0$, the negative binomial distribution has $\text{var}(Y) \rightarrow \mu$ and it converges to the Poisson distribution with mean μ .

The negative binomial has much greater scope than the Poisson. For example, the Poisson mode is the integer part of the mean and thus equals 0 only when $\mu < 1$. The negative binomial mode is the integer part of $\mu(k-1)/k$ (Johnson et al. 2005, p. 217) and can be 0 for any μ .

For independent observations from a negative binomial distribution, the ML estimate of μ is the sample mean, but ML estimation for γ requires iterative methods (R. A. Fisher showed this in an appendix of a 1953 *Biometrics* article by C. Bliss). An alternative gamma parameterization implies a linear rather than quadratic variance function for the negative binomial (Exercise 14.29).

14.4.2 Negative Binomial Regression Modeling

Negative binomial models for counts permit μ to depend on explanatory variables. Such models normally take γ to be the same for all observations. This corresponds to a constant coefficient of variation in the gamma mixing distribution, $\sqrt{\text{var}(\lambda)}/\text{E}(\lambda) = \sqrt{\gamma}$, with the standard deviation increasing as the mean does. Most common is the log link, as in Poisson loglinear models. Sometimes the identity link is adequate, such as with a single predictor that is a factor.

For γ fixed, a negative binomial model is a GLM. The likelihood equations for the regression parameters β are then special cases of those [see (4.25)] for an ordinary GLM with variance function $v(\mu) = \mu + \gamma\mu^2$. The usual iterative reweighted least-squares algorithm applies for ML model fitting. The full log likelihood $L(\beta, \gamma; y)$ for a negative binomial model with link function g satisfies

$$\frac{\partial^2 L}{\partial \beta_j \partial \gamma} = - \sum_i \frac{y_i - \mu_i}{(1 + \gamma \mu_i)^2 g'(\mu_i)} x_{ij}.$$

Thus, $E(\partial^2 L / \partial \beta_j \partial \gamma) = 0$ for each j . Similarly, the inverse of the expected information matrix has 0 elements connecting γ with each β_j . Since this is the asymptotic covariance matrix, $\hat{\beta}$ and $\hat{\gamma}$ are asymptotically independent.

14.4.3 Example: Frequency of Knowing Homicide Victims

[Table 14.6](#) summarizes responses of 1308 subjects to the question: Within the past 12 months, how many people have you known personally that were victims of homicide? The table shows responses by race, for those who identified their race as white or as black. The sample mean for the 159 blacks was 0.522, with a variance of 1.150. The sample mean for the 1149 whites was 0.092, with a variance of 0.155.

Table 14.6 Number of Victims of Murder Known in Past Year, by Race, with Fit of Poisson and Negative Binomial Models

Response	Data		Poisson GLM		Neg. Bin. GLM		Poisson GLMM	
	Black	White	Black	White	Black	White	Black	White
0	119	1070	94.3	1047.7	122.8	1064.9	116.7	1068.3
1	16	60	49.2	96.7	17.9	67.5	24.5	65.3
2	12	14	12.9	4.5	7.8	12.7	8.1	10.1
3	7	4	2.2	0.1	4.1	2.9	3.6	2.8
4	3	0	0.3	0.0	2.4	0.7	1.9	1.1
5	2	0	0.0	0.0	1.4	0.2	1.1	0.5
6	0	1	0.0	0.0	0.9	0.1	0.7	0.3

Source: 1990 General Social Survey.

A natural first choice for modeling count data is a Poisson GLM, such as a loglinear model with an indicator predictor for race. Let y_{it} denote the response for subject t of race i . For $\mu_{it} = E(Y_{it})$, this model is

$$\log \mu_{it} = \alpha + \beta x_{it},$$

with $x_{1t} = 1$ (blacks) and $x_{2t} = 0$ (whites). This model has fit $\hat{\mu}_{it} = -2.38 + 1.733x_{it}$. The estimated expected responses are $\exp(-2.38 + 1.733) = 0.522$ for blacks and $\exp(-2.38) = 0.092$ for whites, the sample means. For any link function for this model, the likelihood equations imply that the fitted means equal the sample means. Since $\beta = 1.733$ ($SE = 0.147$) is the difference between the log means for blacks and whites, the ratio of sample means is $\exp(1.733) = 5.7 = 0.522/0.092$. [Table 14.6](#) also shows the fit of this model.

However, the data show evidence of overdispersion for a Poisson GLM, as for each race the sample variance is roughly double the mean. This evidence is reflected by the higher observed counts at $y = 0$ and at large y values than the Poisson GLM predicts. A negative binomial mixture model seems plausible. Due to demographic factors, heterogeneity probably occurs among subjects of a given race in the distribution of Y . For ML fitting, the deviance decreases by 122.2 compared with the ordinary Poisson GLM that is the special case with $\gamma = 0$. [Table 14.6](#) also shows this model fit. It is dramatically better at $y = 0$ and 1.

[Table 14.7](#) shows parameter estimates for the negative binomial and Poisson GLMs. For both, $\beta = 1.733$ since both models provide fitted means equal to the sample means. However, the estimated standard error of β increases from 0.147 for the Poisson GLM to 0.238 for the negative binomial GLM. The Wald 95% confidence interval for the ratio of means for blacks and whites is $\exp[1.733 \pm 1.96(0.147)] = (4.2, 7.5)$ for the Poisson GLM but $\exp[1.733 \pm 1.96(0.238)] = (3.5, 9.0)$ for the negative binomial GLM. In accounting for the overdispersion, we obtain results that are not as precise as the more naive model suggests but are more credible.

Table 14.7 Parameter Estimates for Models Fitted to Homicide Data

Term	Models with Log Link			Models with Identity Link	
	Neg. Binomial GLM	Poisson GLM	Poisson GLMM	Neg. Binomial GLM	Poisson GLM
α	-2.38	-2.38	-3.69	0.092	0.092
β	1.733	1.733	1.897	0.430	0.430
$SE(\hat{\beta})$	0.238	0.147	0.246	0.109	0.058

The negative binomial model has $\hat{\gamma} = 4.94$ ($SE = 1.00$). This shows strong evidence that $\gamma > 0$, indicating that the negative binomial model is more appropriate than the Poisson GLM. The estimated variance of Y is $\mu C + \hat{\gamma}\mu^2 = \hat{\mu} + 4.94\hat{\mu}^2$, which is 0.13 for whites and 1.87 for blacks, much closer to the sample values than the Poisson model provides.

[Table 14.7](#) also shows results for negative binomial and Poisson models using the identity link. Again, the fits reproduce the sample means and are more imprecise but more credible with the negative binomial model. For this link also the estimated dispersion parameter is $\hat{\gamma} = 4.94$.

14.5 POISSON REGRESSION WITH RANDOM EFFECTS

We've seen that a flexible way to account for overdispersion is with a mixture model. We've just seen that mixing the Poisson using the gamma distribution yields the negative binomial marginally. An alternative mixes the Poisson log mean with a normal random effect.

14.5.1 A Poisson GLMM

Breslow (1984) and Hinde (1982) suggested the GLMM structure (13.1) with the log link and normal random intercept. The model for the mean for observation t in cluster i is

$$(14.14) \log[E(Y_{it}|u_i)] = \mathbf{x}_{it}^T \boldsymbol{\beta} + u_i,$$

where $\{u_i\}$ are independent $N(0, \sigma^2)$. Conditional on u_i , y_{it} has a Poisson distribution. Marginally, the distribution has variance greater than the mean whenever $\sigma > 0$. The identity link is also possible but has a structural problem: When $\sigma > 0$, a positive probability exists that the linear predictor is negative.

The negative binomial model (for fixed γ) is a GLMM with nonnormal random effect. With the log link, it results from a loglinear model of form (14.14) with random intercept, where $\exp(u_i)$ has a gamma distribution with mean 1 and variance γ .

14.5.2 Marginal Model Implied by Poisson GLMM

The Poisson GLMM (14.14) implies a relatively simple marginal model, averaging out the random effect. The mean of the marginal distribution is

$$E(Y_{it}) = E[E(Y_{it}|u_i)] = E[e^{x_{it}^T \beta + u_i}] = e^{x_{it}^T \beta + \sigma^2/2}.$$

Here $E[\exp(u_i)] = \exp(\sigma^2/2)$ because a $N(0, \sigma^2)$ variate u_i has moment generating function $E[\exp(tu_i)] = \exp(t\sigma^2/2)$. So, for the Poisson GLMM the log of the mean conditionally equals $x_{it}^T \beta + u_i$ and marginally equals $x_{it}^T \beta + \sigma^2/2$. A loglinear model still applies. The marginal effects of the explanatory variables are the same as the cluster-specific effects. Thus, the *ratio* of means at two different settings of x_{it} is the same conditionally and marginally. However, marginally the intercept is offset. (Note that Jensen's inequality applies, since the link is not linear.)

The variance of the marginal distribution is

$$\begin{aligned} \text{var}(Y_{it}) &= E[\text{var}(Y_{it}|u_i)] + \text{var}[E(Y_{it}|u_i)] = E[e^{x_{it}^T \beta + u_i}] + e^{2x_{it}^T \beta} \text{var}(e^{u_i}) \\ &= e^{x_{it}^T \beta + \sigma^2/2} + e^{2x_{it}^T \beta} (e^{2\sigma^2} - e^{\sigma^2}) = E(Y_{it}) + [E(Y_{it})]^2 (e^{\sigma^2} - 1). \end{aligned}$$

Here, $\text{var}(e^{u_i}) = E(e^{2u_i}) - [E(e^{u_i})]^2 = e^{2\sigma^2} - e^{\sigma^2}$ by evaluating the moment generating function at $t = 2$ and $t = 1$. As in the negative binomial model, the marginal variance is a quadratic function of the marginal mean. It exceeds the marginal mean when $\sigma > 0$. The ordinary Poisson model results when $\sigma = 0$. When $\sigma > 0$ the marginal distribution is not Poisson, and the extent to which the variance exceeds the mean increases as σ increases.

As in binary GLMMs, Y_{it} and Y_{is} are independent given u_i but are marginally nonnegatively correlated. For $t \neq s$,

$$\begin{aligned} \text{cov}(Y_{it}, Y_{is}) &= E[\text{cov}(Y_{it}, Y_{is}|u_i)] + \text{cov}[E(Y_{it}|u_i), E(Y_{is}|u_i)] \\ (14.15) \quad &= 0 + \text{cov}[\exp(x_{it}^T \beta + u_i), \exp(x_{is}^T \beta + u_i)]. \end{aligned}$$

The functions in the last covariance term are both monotone increasing functions of u_i , and hence are nonnegatively correlated (Exercise 14.33).

14.5.3 Example: Homicide Victim Frequency Revisited

For [Table 14.6](#) on responses of the number of known victims of homicide within the past 12 months, models permitting subject heterogeneity are sensible. For the response y_{it} for subject t of race i , the Poisson GLMM is

$$\log[E(Y_{it}|u_{it})] = \alpha + \beta x_{it} + u_{it},$$

where $\{u_{it}\}$ are independent $N(0, \sigma^2)$. The log means vary according to a $N(\alpha, \sigma^2)$ distribution for whites and a $N(\alpha + \beta, \sigma^2)$ distribution for blacks. Given u_{it} , y_{it} has a Poisson distribution.

[Table 14.6](#) also shows this model fit, and [Table 14.7](#) shows estimates. The random effects have $\hat{\sigma} = 1.63$ ($SE = 0.15$). The deviance decreases by 116.6 compared with the Poisson GLM, indicating a better fit by allowing heterogeneity. For subjects at the means of the random effects distributions ($u_{it} = 0$) the estimated expected responses are $\exp(-3.69 + 1.90) = 0.167$ for blacks and $\exp(-3.69) = 0.025$ for whites. The fitted marginal mean is $\exp(\hat{\alpha} + \hat{\beta}x_{it} + \hat{\sigma}^2/2)$, or 0.63 for blacks and 0.09 for whites. The fitted marginal variances are 0.21 for blacks and 5.78 for whites. These are somewhat larger than the sample means and variances, perhaps because the fitted distribution has nonnegligible mass above the largest observed response of 6.

14.5.4 Negative Binomial Models versus Poisson GLMMs

The Poisson GLMM with normal random effects has the advantage, relative to the negative binomial GLM, of easily permitting multivariate random effects and multilevel models. However, the negative binomial has properties that can make interpretation simpler. We've seen that the identity link is valid for it, which is useful for simple examples such as the preceding one with a factor predictor. With any link and a factor predictor, its ML fitted means equal the sample means. This is not the case for the Poisson GLMM.

NOTES

Section 14.1: Latent Class Models

14.1 Latent variables: For fitting and interpretation of latent class and related latent variable models, see Aitkin et al. (1981), Bartholomew et al. (2011), Clogg (1995), Clogg and Goodman (1984), Collins and Lanza (2009), Goodman (1974), Haberman (1979, Chap. 10), Hagenaars and McCutcheon (2009), Heinen (1996), Lazarsfeld and Henry (1968), Magidson and Vermunt (2004), Skrondal and Rabe-Hesketh (2004), and Vermunt (2003). For a similar *mixed-membership model*, each subject has partial membership in various classes, with a distribution specifying a probability for membership in a class (Erosheva et al. 2007). Espeland and Handelman (1989), Uebersax (1993), Uebersax and Grove (1990, 1993), and Yang and Becker (1997) presented latent variable models for rater agreement and diagnostic accuracy.

14.2 Mixture goodness of fit: Rudas et al. (1994) proposed a clever mixture method for summarizing goodness of fit. For a model M for a contingency table with true probabilities π , they used the mixture $\pi = (1 - \rho)\pi_1 + \rho\pi_2$, with π_1 the model-based probabilities and π_2 unconstrained. Their index of lack of fit is the smallest such ρ possible for which this holds. It is the fraction of the population that cannot be described by the model. This recognizes that any given model does not truly hold but is useful if ρ is close to 0.

Section 14.2: Nonparametric Random Effects Models

14.3 Rasch and QS: For connections between Rasch-type models and quasi-symmetry models, see Agresti (1993, 1997), Conaway (1989), Darroch (1981), Darroch et al. (1993), and Kelderman (1984).

Section 14.3: Beta-Binomial Models

14.4 Beta-binomial references: Skellam (1948) introduced the beta-binomial distribution. For modeling using this distribution or related quasi-likelihood approaches, see Albert (2010), Brooks et al. (1997), Capitanu and Presnell (2008), Crowder (1978), Hinde and Demétrio (1998), Lee et al. (2006), Liang and Hanfelt (1994), Liang and McCullagh (1993), Lindsey and Altham (1998), Moore (1986a), Moore and Tsiatis (1991), Nelder and Pregibon (1987), Prentice (1986), Rosner (1984, 1989) [with critique by Neuhaus and Jewell (1990a)], Slaton et al. (2000), and Williams (1975, 1982). For beta-binomial type variance, Ryan (1995) and Williams (1988) showed advantages of the quasi-likelihood approach over ML. The beta-binomial generalizes to a Dirichlet-multinomial: Conditional on the probabilities, the distribution is multinomial, and the probabilities themselves have a Dirichlet distribution. See Brier (1980), Guimarães (2005), Guimarães and Lindrooth (2007), Mosimann (1962), Paul et al. (1989), and Exercise 14.30.

14.5 Developmental toxicity: For modeling overdispersion caused by litter effects in developmental toxicity studies with binary data, see Follman and Lambert (1989), Kupper and Haseman (1978), Kupper et al. (1986), Lefkopoulos et al. (1989), and Ryan (1992). Ochi and Prentice (1984) proposed a probit model based on an underlying normal latent variable model with common pairwise correlations.

Section 14.4: Negative Binomial Regression

14.6 NB modeling: Johnson et al. (2005, Chap. 5) summarized properties of the negative binomial distribution. Cameron and Trivedi (1998, p. 72) showed the asymptotic covariance matrix of model parameter estimates. They and Lawless (1987) considered a moment estimator for γ and studied robustness properties. They noted that $\hat{\beta}$ is consistent if the model for the mean is correctly specified, even if the true distribution is not negative binomial. Booth et al. (2003),

Hilbe (2011), and Hinde and Demétrio (1998) also discussed NB modeling.

Section 14.5: Poisson Regression with Random Effects

14.7 Zero-inflated models: Overdispersion relative to the Poisson distribution often occurs when the frequency of 0 outcomes is larger than expected. One way to deal with this is a mixture model that mixes a distribution that is degenerate at 0 with an ordinary Poisson (or negative binomial) distribution. See Min and Agresti (2005) for details and references.

EXERCISES

Applications

14.1 Create a 2^5 table of opinions about legalized abortion by downloading the table for the items labeled (ABRAPE, ABHLTH, ABSINGLE, ABDEFECT, ABPOOR) in the most recent GSS. Fit a latent class model. For each latent class, find the estimated probability of supporting legalized abortion the five situations. Suggest a tentative interpretation for the classes.

14.2 Fit a logistic-normal random effects model to the carcinoma ratings of [Table 14.1](#). Compare results to those for latent class models in Section 14.1.3.

14.3 For capture–recapture experiments, Coull and Agresti (1999) used a quasi-symmetric loglinear model with no higher-order terms,

$$\log \mu(y_1, \dots, y_T) = \lambda + \beta_1 y_1 + \dots + \beta_T y_T + \beta(y_1 y_2 + y_1 y_3 + \dots + y_{T-1} y_T).$$

Show that **(a)** like the logistic-normal GLMM, this model has exchangeable association and only one more parameter than the mutual independence model, **(b)** the fit to [Table 13.6](#) yields $\hat{N} = 90.5$ and a 95% profile-likelihood confidence interval for N of (75, 125).

14.4 A data set on pregnancy rates among girls under 18 years of age in 13 north central Florida counties has information on a 3-year total for each county i on n_i = number of births and y_i = number of those for which the mother's age was under 18 (see J. Booth, in *Statistical Modelling: Lecture Notes in Statistics*, 104, Springer, 43–52, 1995).

- a.** For a beta-binomial model, the ML estimated parameters are $\hat{\alpha}_1 = 9.9$ and $\hat{\alpha}_2 = 240.8$. Use the mean and variance to describe the estimated beta distribution and the estimated marginal distribution of Y_i (as a function of n_i).
- b.** Quasi-likelihood using variance function [\(14.10\)](#) for the model $\text{logit}(\mu_i) = \alpha$ has $\hat{\alpha} = -3.18$ and $\hat{\rho} = 0.005$. Describe the estimated mean and variance of Y_i .
- c.** Quasi-likelihood using variance [\(14.11\)](#) for the model $\text{logit}(\mu_i) = \alpha$ has $\hat{\alpha} = -3.35$ and $\hat{\phi} = 8.3$. Describe the estimated mean and variance of Y_i .
- d.** The logistic-normal GLMM, $\text{logit}(\pi_i) = \alpha + u_i$, yields $\hat{\alpha} = -3.24$ and $\hat{\sigma} = 0.33$. Describe the estimated mean of Y_i [Recall [\(13.9\)](#)].

14.5 In Exercise 13.2 about Ray Allen's three-point shooting, the simple binomial model, $\pi_i = \alpha$, has lack of fit. Fit the beta-binomial model or use the quasi-likelihood approach with that variance structure. Use the fit to summarize his free-throw shooting, by giving an estimated mean and standard deviation for π_i .

14.6 Extend the various analyses of the teratology data in Section 14.5 as follows:

- a.** Include a predictor for litter size (as well as group). Interpret, and compare results to those without this predictor.
- b.** Fit a model with beta-binomial variance [\(14.10\)](#) in which ρ varies by treatment group. Use results to motivate a model that allows overdispersion only in the placebo group. Interpret and compare results to those with common ρ for each group.

14.7 [Table 14.8](#) reports the results of a study of fish hatching under three environments. Eggs from seven clutches were randomly assigned to three treatments, and the response was whether an egg hatched by day 10. The three treatments were (1) carbon dioxide and oxygen removed, (2) carbon dioxide only removed, and (3) neither removed.

[Table 14.8](#) Data for Exercise 14.7

Clutch	Treatment 1		Treatment 2		Treatment 3	
	Number Hatched	Total	Number Hatched	Total	Number Hatched	Total
1	0	6	3	6	0	6
2	0	13	0	13	0	13
3	0	10	8	10	6	9
4	0	16	10	16	9	16
5	0	32	25	28	23	30
6	0	7	7	7	5	7
7	0	21	10	20	4	20

Source: Data courtesy of Becca Hale, Zoology Department, University of Florida.

- a. Let π_{it} denote the probability of hatching for an egg from clutch i in treatment t . Assuming independent binomial observations, fit the model

$$\text{logit}(\pi_{it}) = \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3,$$

where $z_t = 1$ for treatment t and 0 otherwise. What does your software report for $\hat{\beta}_1$, and what should it be? [Hint: Note that treatment 1 has no successes.]

- b. Analyze these data using an approach that allows overdispersion. Interpret. Indicate whether evidence of overdispersion occurs for treatments 2 and 3.

14.8 Copy the “Ohio Children Wheeze Status” data at the website cran.r-project.org/web/packages/geepack/geepack.pdf for the *geepack* package in R. Analyze these data using one method from each of Chapters 12, 13, and 14. Compare results and interpret.

14.9 In 2002 the General Social Survey asked “How many people at your work place are close friends?” The 756 responses had a mean of 2.76, standard deviation of 3.65, and a mode of 0. If you plan to build a GLM using some explanatory variables for this response, which distribution might be sensible? Why?

14.10 One question in a GSS asked subjects how many times they had sexual intercourse in the preceding month.

- a. The sample means were 5.9 for males and 4.3 for females; the sample variances were 54.8 and 34.4. The mode for each gender was 0. Does an ordinary Poisson GLM seem appropriate? Explain.

- b. The Poisson GLM with log link and an indicator variable for gender (1 = males, 0 = females) has gender estimate 0.308 ($SE = 0.038$). Find the Wald 95% confidence interval for the ratio of means for males and females.

- c. For the negative binomial model, the log likelihood increases by 248.7. The estimated difference between the log means is also 0.308, but now $SE = 0.127$. Find the 95% confidence interval for the ratio of means. Compare to the Poisson GLM, and interpret. Which do you think is more appropriate? Why?

14.11 For the data in the previous exercise, argue that a possibly more realistic model assumes for gender i a proportion ρ_i that is necessarily 0 and a proportion $1 - \rho_i$ that has distribution that is a gamma mixture of Poissons.

14.12 For the homicide data, reproduce the results in [Table 14.7](#) for the identity link. Explain why the estimated difference in means is identical for the two GLMs but the SE values are very different. Use the more appropriate one to form a confidence interval for the true difference in means.

14.13 For the horseshoe crab satellite counts in [Table 4.3](#), use width as a predictor.

- a. Fit a negative binomial model with log link. Interpret. Describe the estimated variance as a function of μ .

- b. Fit a Poisson GLMM with log link. Interpret.

- c. Compare results for the models, including those in Section 4.3.2 for Poisson and negative binomial GLMs. Indicate your preferred model. Justify.

14.14 Use quasi-likelihood methods to analyze [Table 14.6](#) on counts of murder victims.

14.15 Refer to Exercise 4.1. With data at the book's website, use methods of this chapter to analyze how the countywide vote for Pat Buchanan in 2000 related to the vote for Ross Perot in 1996. Note that Palm Beach County is an enormous outlier. Model with and without that observation and compare results.

Theory and Methods

14.16 When $I = 2$, for $q \geq 2$ show that we need $T \geq 4$ for the latent class model to be unsaturated. Then, find the maximum value for q when $T = 4, 5$. For an I^2 table, show we need $q < I^2/(2I - 1)$.

14.17 Express the log likelihood for latent class model (14.1) in terms of the model parameters. Derive likelihood equations (Goodman 1974, Haberman 1979).

14.18 In Section 14.2.3, under the null that the ordinary logistic regression model holds, explain why it is inappropriate to treat the difference between the deviances for that model and the mixture of two logistic regressions as a chi-squared statistic.

14.19 Express the numerator of the beta density in terms of μ and θ . Using this, show that it is (a) unimodal when $\theta < \min(\mu, 1 - \mu)$, and (b) the uniform density when $\mu = \theta = \frac{1}{2}$.

14.20 Suppose $\text{corr}(Y_{it}, Y_{is}) = \rho$ for $t \neq s$. Show that $\text{var}(Y_{it}) = \pi_i(1 - \pi_i)$, $\text{cov}(Y_{it}, Y_{is}) = \rho\pi_i(1 - \pi_i)$, and

$$\text{var}\left(\sum_t Y_{it}\right) = n_i\pi_i(1 - \pi_i)[1 + \rho(n_i - 1)].$$

14.21 Show that the beta-binomial distribution (14.9) simplifies to the binomial when (a) $\theta = 0$, (b) $n = 1$. Explain why overdispersion cannot occur when $n = 1$.

14.22 Liang and Hanfelt (1994) described a teratology study comparing control and treatment groups in which the ML estimate of the treatment effect in a beta-binomial model differs by a factor of 2 depending on whether you assume the same overdispersion parameter for each group. By contrast, with variance function (14.11), the quasi-likelihood estimate of the treatment effect is the same whether you assume the same or different ϕ for the two groups. Explain why, and discuss whether this is an advantage or disadvantage of that method.

14.23 For small σ , show that the logistic-normal model, $\text{logit}(\pi_i) = \alpha + \mathbf{x}_i^T \boldsymbol{\beta} + u_i$, corresponds approximately to a mixture model for which the mixture distribution has $\text{var}(\pi_i) = [\mu_i(1 - \mu_i)]^2\sigma^2$. [Hint: See Exercise 4.35.]

14.24 Altham (1978) introduced the discrete distribution

$$f(y; \pi, \psi) = c(\pi, \psi) \binom{n}{y} \pi^y (1 - \pi)^{n-y} \exp[\psi y(n - y)], \quad y = 0, 1, \dots, n,$$

where $c(\pi, \psi)$ is a normalizing constant. Show that this is in the exponential family. Show that the binomial occurs when $\psi = 0$. [Altham noted that overdispersion occurs when $\psi < 0$. Corcoran et al. (2001) and Lindsey and Altham (1998) used this as the basis of an alternative model to the beta-binomial.]

14.25 Refer to the previous exercise. For n identically distributed but correlated binary observations (y_1, y_2, \dots, y_n) , a related loglinear model is

$$\log \mu(y_1, y_2, \dots, y_n) = \lambda \left(\sum_i y_i \right) + \nu \left(\sum_{h < i} \sum y_h y_i \right).$$

Explain why this is a simple special case of the quasi-symmetry model, and explain how the binomial is a special case.

14.26 When y_1, \dots, y_N are independent from a negative binomial distribution (14.13) with γ fixed, show that $\bar{\mu} = \bar{y}$.

14.27 Using $E(Y) = E[E(Y|X)]$ and $\text{var}(Y) = E[\text{var}(Y|X)] + \text{var}[E(Y|X)]$, derive the mean and variance of the (a) beta-binomial distribution, and (b) negative binomial distribution.

14.28 Suppose that given u , Y is Poisson with $E(Y|u) = u\mu$, where μ may depend on predictors. Suppose that u is a positive random variable with $E(u) = 1$ and $\text{var}(u) = \tau$. Show that $E(Y) = \mu$

and $\text{var}(Y) = \mu + \tau\mu^2$. Explain how negative binomial GLMs and Poisson GLMMs with log link can follow as special cases.

14.29 An alternative negative binomial parameterization results from the gamma density formula,

$$f(\lambda; k, \mu) = \frac{k^{k\mu}}{\Gamma(k\mu)} \exp(-k\lambda) \lambda^{k\mu-1}, \quad \lambda \geq 0,$$

for which $E(\lambda) = \mu$, $\text{var}(\lambda) = \mu/k$. Show that this gamma mixture of Poissons yields a negative binomial with

$$E(Y) = \mu, \quad \text{var}(Y) = \mu(1+k)/k.$$

For what limiting value of k does this reduce to the Poisson? [See Lee and Nelder (1996) for ML model fitting. Cameron and Trivedi (1998, p. 75) pointed out that, unlike with quadratic variance, consistency does not occur for the GLM parameter estimators when the model for the mean holds but the true distribution is not negative binomial.]

14.30 Suppose Y_1 and Y_2 are independent negative binomial variates with common dispersion parameter γ . Show that $Y_1 + Y_2$ is negative binomial with dispersion parameter $\gamma/2$. Show that Y_1 , conditional on $Y_1 + Y_2$, is beta-binomial. State the multiple-category extension that yields a *Dirichlet-multinomial* distribution. Explain the analogy with the Poisson-multinomial result in Section 1.2.5.

14.31 Show that the loglinear random effects model

$$\log[E(Y_{it}|\boldsymbol{u}_i)] = \mathbf{x}_{it}^T \boldsymbol{\beta} + \mathbf{z}_{it}^T \boldsymbol{u}_i,$$

where $\{\boldsymbol{u}_i\}$ are independent $N(\mathbf{0}, \Sigma)$, implies the marginal loglinear model

$$\log[E(Y_{it})] - \frac{1}{2}\mathbf{z}_{it}^T \boldsymbol{\Sigma} \mathbf{z}_{it} = \mathbf{x}_{it}^T \boldsymbol{\beta},$$

with the same fixed effects but with offset term.

14.32 In Section 14.5.2 and the previous exercise we saw that for Poisson GLMMs, the marginal effects are the same as the cluster-specific effects. This does not imply that ML estimates of effects are the same for a Poisson GLMM and a Poisson GLM. Explain why. [Hint: For the GLMM, is the marginal distribution Poisson?]

14.33 For the Poisson GLMM (14.14), use the normal moment generating function to show that, for $t \neq s$,

$$\text{cov}(Y_{it}, Y_{is}) = \exp[(\mathbf{x}_{it}^T + \mathbf{x}_{is}^T)\boldsymbol{\beta}] [\exp(\sigma^2)(\exp(\sigma^2) - 1)].$$

Hence, find $\text{corr}(Y_{it}, Y_{is})$.

14.34 For a Poisson GLMM using the identity link, relate the marginal mean and variance to the conditional mean and variance. Explain the structural problem that this model has.

CHAPTER 15

Non-Model-Based Classification and Clustering

In this book we've focused on ways of *modeling* categorical response data. This chapter presents some alternative analyses that are not model-based or else have a much more general model structure.

Sections 15.1 and 15.2 deal with non-model-based alternatives to logistic regression for classifying observations into response categories. In Section 15.1 we introduce *linear discriminant analysis*, a method that is more efficient than logistic regression when the explanatory variables have a normal distribution. In Section 15.2 we present a method for constructing a graphical tree for making such predictions. In Section 15.3 we discuss ways of grouping sets of observations on multiple response variables into clusters.

15.1 CLASSIFICATION: LINEAR DISCRIMINANT ANALYSIS

In Section 6.3.3 we used logistic regression to classify binary observations. One rule predicts that $y = 1$ whenever the \mathbf{x} values are such that the model has an estimate $\hat{\pi}$ of $P(y = 1)$ that exceeds 0.50. Equivalently, this corresponds to having linear predictor value (including the intercept) in the model satisfy $\hat{\beta}^T \mathbf{x} > 0$. There are alternative, non-model-based ways of dividing the set of explanatory variable values into two sets, in one of which the predicted $y = 1$ (which we denote by $\hat{y} = 1$) and in the other of which $\hat{y} = 0$.

We've seen one such method in Section 7.4.4, using the kernel approach of nearest neighbor smoothing. The best known non-model-based method, called *linear discriminant analysis*, is another simple alternative to logistic regression for binary classification. Recall that logistic regression makes no assumption about the distribution of \mathbf{X} and instead focuses on the binary distribution of Y given \mathbf{x} . By contrast, linear discriminant analysis also makes an assumption about the distribution of \mathbf{X} , given y . For it, like logistic regression, the boundary between the two sets of \mathbf{x} values with $\hat{y} = 1$ and $\hat{y} = 0$ is linear. Why do this instead of logistic regression? When the normality assumption is reasonable, there is the potential of an efficiency improvement from using the extra information.

15.1.1 Classification with Normally Distributed Predictors

In Section 5.1.5 (and Exercise 5.30) we noted that normal distributions for $(X|Y=j)$ for $j = 0, 1$ imply a logistic regression curve for $P(Y=1|x)$. For multiple predictors, suppose that $(X|Y=j)$ has a multivariate $N(\boldsymbol{\mu}_j, \Sigma)$ distribution, $j = 0, 1$. Then, by Bayes' theorem with $\pi = P(Y=1)$, it follows that $P(Y=1|x)$ satisfies

$$\text{logit}[P(Y=1|x)] = \log \frac{\pi}{1-\pi} - \frac{1}{2}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1}x.$$

That is, a logistic regression model holds with effect parameters $\boldsymbol{\beta} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1}$. The effects are stronger when the groups having $y = 1$ and having $y = 0$ are farther apart and when there is less variability within those groups.

Fisher (1936) developed a related method for using observations on x to classify on y , before the advent of logistic regression. It assumes a common covariance matrix Σ for X within each category for y , but its motivation does not require normality assumptions. Fisher's goal was to find a linear combination $\ell^T x$ such that its values when $y = 1$ were separated as much as possible from its values when $y = 0$, relative to the variability of $\ell^T x$ values within each y category. The solution maximizes the squared distance between the means of $\ell^T x$ for the two categories of y , divided by the within-category variance of $\ell^T x$. Equivalently, for a given value of x , the prediction for y is the category j ($j = 0, 1$) that has the minimum of the Mahalanobis distance of x from $\boldsymbol{\mu}_j = E(X|Y=j)$, which is

$$d_j(x) = (x - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1}(x - \boldsymbol{\mu}_j), \quad j = 0, 1.$$

When we have a prior value π_0 for $P(Y=1)$, then $\text{logit}(\pi_0)$ is subtracted from the distance for $j = 1$.

In practice, we estimate these distances by substituting the sample means \bar{x}_1 and \bar{x}_0 and a pooled covariance estimate S for Σ . This method yields a linear function as the boundary between the sets of x values having $\hat{y} = 1$ and $\hat{y} = 0$. At a particular x , $\hat{y} = 1$ if

$$(\bar{x}_1 - \bar{x}_0)^T S^{-1} x > (\bar{x}_1 - \bar{x}_0)^T S^{-1}(\bar{x}_1 + \bar{x}_0)/2 - \text{logit}(\pi_0).$$

For example, with $\pi_0 = 0.50$ and a single predictor x having $\bar{x}_1 > \bar{x}_0$, we predict that $y = 1$ if $x > (\bar{x}_1 + \bar{x}_0)/2$, that is, if x is closer to \bar{x}_1 than to \bar{x}_0 .

This prediction rule depends on x only through the left-hand term in this equation. This term, $(\bar{x}_1 - \bar{x}_0)^T S^{-1} x$, is called *Fisher's linear discriminant function*. In fact, the regression function for ordinary least-squares regression of an indicator variable for y on x (which is the ML fit of the linear probability model under a normal response assumption) is proportional to that term. Because of this connection between Fisher's linear discriminant function and the regression equation, the observations having $\hat{y} = 1$ are those for which the linear regression-based estimate of $E(Y|x)$ is sufficiently high.

Consider now the additional assumption that the distribution of X in each y category is multivariate normal with common covariance matrix. Then, Bayes' theorem with a particular prior value $\pi_0 = P(Y=1)$ provides proper posterior probability estimates for each category. With $\pi_0 = 0.50$ and the estimated Mahalanobis distance values, at a particular value of x ,

$$\hat{P}(Y=1|x) = \frac{\exp[-\frac{1}{2}\hat{d}_1^2(x)]}{\exp[-\frac{1}{2}\hat{d}_0^2(x)] + \exp[-\frac{1}{2}\hat{d}_1^2(x)]}.$$

Section 6.3.3 presented the *classification table* as a way of summarizing predictions made using a fitted logistic regression model. This type of table can describe the quality of predictions with any method of classification. The true misclassification probabilities tend to be underestimated by predicting observations using the equation to which those observations contributed, so cross-validation can be employed to obtain less biased estimates.

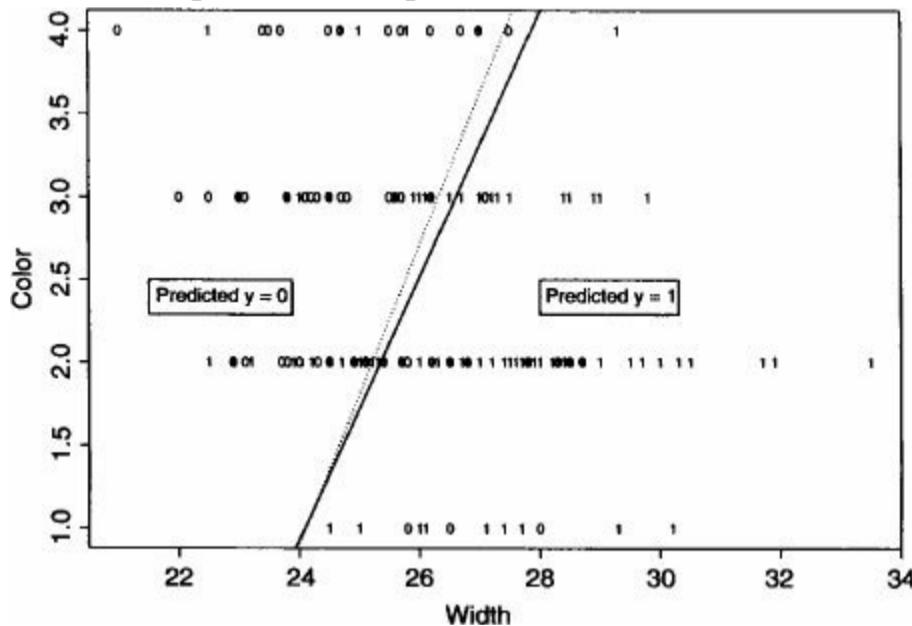
15.1.2 Example: Horseshoe Crab Satellites Revisited

In Section 6.3.3 we illustrated classification tables for logistic regression using the horseshoe crab data set, for the model using a female crab's width and color as predictors of whether that crab has at least one male satellite. To illustrate linear discriminant analysis, as in model (5.14) in Section 5.4.6 we'll use the quantitative scoring (1,2,3,4) for the color levels rather than treating color as a factor, so that the normality assumption for the joint distribution of $x = \text{width}$ and $c = \text{color}$ is not so badly violated. For $\pi_0 = 0.50$, software (SAS PROC DISCRIM) reports the linear discriminant function $0.430x - 0.553c$, with $\hat{y} = 1$ when $0.430x - 0.553c > 9.811$.

The least-squares fit of the linear probability model for y is $\hat{\pi} = -1.234 + 0.0811x - 0.1024c$. The coefficients of x and c are identical to those from the linear discriminant function divided by 5.30. The inequality for predicting y based on the linear discriminant function (i.e., $\hat{y} = 1$ when $0.430x - 0.553c > 9.811$) is equivalent to $-1.851 + 0.0811x - 0.1024c > 0$, or $\hat{\pi} > 0.614$ for the fit of the linear probability model. (The inequality is equivalent to $\hat{\pi} > 0.50$ only when the sample proportion of $y = 1$ is 0.50.)

[Figure 15.1](#) shows the data and the classification regions obtained with linear discriminant analysis. The boundary line is $c = -18.08 + 0.79x$. The figure also shows the boundary line from using logistic regression with a cutoff of $\hat{\pi} > 0.614$ for $\hat{y} = 1$, for which $c = -20.70 + 0.90x$. In practical terms, the regions are very similar.

[Figure 15.1](#) Classification regions (solid line for linear discriminant analysis, dotted line for logistic regression) with width and color predictors of presence of horseshoe crab satellites.



[Table 15.1](#) shows the classification table that results using cross-validation, in which to predict observation i , we use the linear discriminant function obtained with the other $n - 1$ observations. [Table 15.1](#) also shows a classification table based on logistic regression modeling, with cross-validation. To enhance comparability of the two approaches, we used 0.614 as the boundary for $\hat{\pi}$ for the predictions.

[Table 15.1](#) Classification Tables for Predictions Using Discriminant Analysis and Logistic Regression for Horseshoe Crab Data

Actual	Discriminant Analysis		Logistic Regression		Total
	$\hat{y} = 1$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 0$	
$y = 1$	72	39	77	34	111
$y = 0$	19	43	21	41	62

15.1.3 Multicategory Classification and Other Versions of Discriminant Analysis

In discriminant analysis, the linear separating boundary in the space of \mathbf{x} values between $y = 1$ and $y = 0$ can be generalized. If we include quadratic and cross-product interaction terms in \mathbf{x} , the boundary becomes quadratic in the space of the original \mathbf{x} variables (Anderson 1975). Interestingly, the normal assumption for the distribution of $(\mathbf{X}|Y=j)$ but with unequal covariance matrices implies a logistic regression model for $P(Y=1)$ that is quadratic in \mathbf{x} . For other generalizations, see Note 15.2.

In the other direction (i.e., simplicity), a *diagonal discriminant analysis* simplification treats the common covariance matrices for $(\mathbf{X}|Y=0)$ and $(\mathbf{X}|Y=1)$ as being diagonal. This seems like an assumption that could be badly violated for most applications; however, as discussed in Section 15.1.4, when the number of predictors is very large, it can result in better classification performance than linear or quadratic discriminant analysis.

Linear discriminant analysis extends directly to multicategory classification. When $Y=j$, denote the pdf of \mathbf{X} by $g_j(\mathbf{x})$, and let $\pi_j = P(Y=j)$, $j = 1, 2, \dots, J$. By Bayes' theorem,

$$P(Y = j|\mathbf{X} = \mathbf{x}) = \frac{\pi_j g_j(\mathbf{x})}{\sum_h \pi_h g_h(\mathbf{x})}.$$

If we assume a particular parametric family for $\{g_j\}$, we can use data to estimate the densities and hence estimate classification probabilities for Y .

The most common way to do this assumes that $(\mathbf{X}|Y=j)$ has a multivariate $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ distribution, $j = 1, 2, \dots, J$, with the same covariance matrix for each group. It can then be shown (Warner 1963) that

$$\log \frac{P(Y = j|\mathbf{x})}{P(Y = J|\mathbf{x})} = \log \frac{\pi_j}{\pi_J} - \frac{1}{2} (\boldsymbol{\mu}_j + \boldsymbol{\mu}_J)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_j - \boldsymbol{\mu}_J) + (\boldsymbol{\mu}_j - \boldsymbol{\mu}_J)^T \boldsymbol{\Sigma}^{-1} \mathbf{x}.$$

That is, a baseline-category logit model holds with effect parameters $\boldsymbol{\beta}_j = (\boldsymbol{\mu}_j - \boldsymbol{\mu}_J)^T \boldsymbol{\Sigma}^{-1}$. After estimating the multivariate normal parameters by the sample means $\{\bar{\mathbf{x}}_j\}$ and a pooled covariance estimate \mathbf{S} , the method predicts that $y=j$ if the *linear discriminant function*

$$\bar{\mathbf{x}}_h^T \mathbf{S}^{-1} \mathbf{x} - \frac{1}{2} \bar{\mathbf{x}}_h^T \mathbf{S}^{-1} \bar{\mathbf{x}}_h + \log \hat{\pi}_h$$

takes maximum value for $h=j$.

As in the binary case, if the assumption were truly satisfied about \mathbf{X} having a normal conditional distribution with common covariance, this classification method would be optimal. To avoid such a strong assumption, we can instead use direct ML fitting with the baseline-category logit model.

15.1.4 Classification Methods for High Dimensions

In classification problems with large p , Bickel and Levina (2004) found that prediction rules that treat those explanatory variables as independent often outperform rules that estimate dependences among them in constructing the classifier. This reflects the difficulty in estimating well the covariance matrix when p is very large. They showed that a *naive Bayes* rule that assumes independence, such as in diagonal discriminant analysis, can greatly outperform ordinary linear discriminant analysis.

An assumption of independence seems stringent and grossly invalid for many applications. However, with very large p , we often expect any particular predictor to be very weakly correlated with most of the other predictors, and a true correlation value is typically closer to 0 than to the ML estimate based on an enormous correlation matrix. For example, Dudoit et al. (2002) found that this method performs better than ordinary linear discriminant analysis for classifying tumors using gene expression data. However, Fan and Fan (2008) noted that even for the independence classification rule, performance can be poor because of the accumulation of noise unless there is some variable reduction.

For classification, another simple alternative to logistic regression and linear discriminant analysis is the *nearest neighbors* method (Section 7.4.2). It classifies an observation based on an estimated probability obtained by averaging response values for nearby observations. A challenge with this method is that with a very large p , the “curse of dimensionality” occurs and a subject may have few or no close neighbors.

A more complex method used with large p is *support vector machines*, presented in Section 15.2.6. Zhang et al. (2006) used this approach together with a SCAD-type penalty for the application of identifying important genes for cancer classification. But as discussed in Section 15.2.6, there is no guarantee that more complex methods will have better performance.

15.1.5 Discriminant Analysis Versus Logistic Regression

When the explanatory variables truly have a normal distribution, conditional on y , discriminant analysis is optimal for classifying observations. In particular, it is more efficient than logistic regression, potentially considerably more as the groups become more widely separated (Efron 1975). This is because it utilizes the information about the distribution of X , which logistic regression ignores.

Often, however, explanatory variables can be far from normally distributed, such as when at least one explanatory variable is qualitative. Also, extreme outliers on x can have a large effect on discriminant analysis (as in ordinary linear regression) but little impact on logistic regression. So, logistic regression is more robust and has broader scope, as it makes no assumption about a distribution for X and merely assumes a binomial distribution for Y at each value of x . Also, logistic regression has the advantage over discriminant analysis of providing direct ways of summarizing effects of explanatory variables, through odds ratios. See Note 15.1 and Section 8.5.2 of McLachlan (2004) and references therein for more discussion of the relative merits of the two approaches.

15.2 CLASSIFICATION: TREE-STRUCTURED PREDICTION

In recent years non-model-based methods have been further developed for predicting response variables using data on a set of explanatory variables. These are examples of methods often referred to with the terms *machine learning* and *data mining*. Rather than relying on a model to summarize effects of explanatory variables on the response variable, such methods are algorithm-driven. Using various criteria, they provide a way of “learning” from the available information on all the variables to estimate the unknown relationship between $E(Y)$ and the explanatory variables \mathbf{x} . This results in an algorithm for making future predictions of y based solely on values of \mathbf{x} . The effectiveness of the algorithm is evaluated by its error rate for future samples.

Even with a model, when n is extremely large, significance tests are less relevant, as statistical significance does not imply practical significance. Inference may not even be relevant because of nonprobability sampling. Some strategies may be useful for prediction of response outcomes even if they have complex structure and do not correspond to understandable models.

A detailed presentation of these algorithmic methods is beyond the scope of this book. In this section, we describe a particular method for binary responses that provides a simple tree-structured depiction of how predictions can be made. Compared with discriminant analysis, this classification method is less restrictive in distributional assumptions and in the form for the predictor decision boundary. However, the methods yield decision boundaries that are highly nonlinear in the space of \mathbf{x} values.

15.2.1 Classification Trees

The *classification tree* method formalizes a decision process that uses a sequential set of questions about the x values to yield a classification prediction for y . A created graphical tree summarizes binary splits on variables at various stages to determine the prediction. This method, proposed by Breiman et al. (1984) and extending earlier work such as by Kass (1980), utilizes classification tables in the process of forming the tree. Its set of x values for which $\hat{y} = 1$ has simple form, consisting of a set of rectangular regions.

For example, consider the prediction of a person's vote for the Democrat or Republican candidate in a U.S. presidential election. Two regions that yield a prediction of voting for the Republican candidate might be (1) everyone who is male and attends religious services at least once a week and has annual income over \$50,000, and (2) everyone who is female and who opposes legalized abortion and is married and never been divorced. A common application of classification trees is making a prediction about whether a patient has a particular medical condition. Zhang and Singer (2010) described an early application that used responses to 13 questions based on results of physiological tests and various patient characteristics (such as age and medical history) to predict whether a patient arriving at an emergency room complaining of chest pain has had a heart attack. Breiman et al. (1984, Chap. 6) showed a similar sort of application, classifying the prognosis of heart attack victims as survivors or as early deaths.

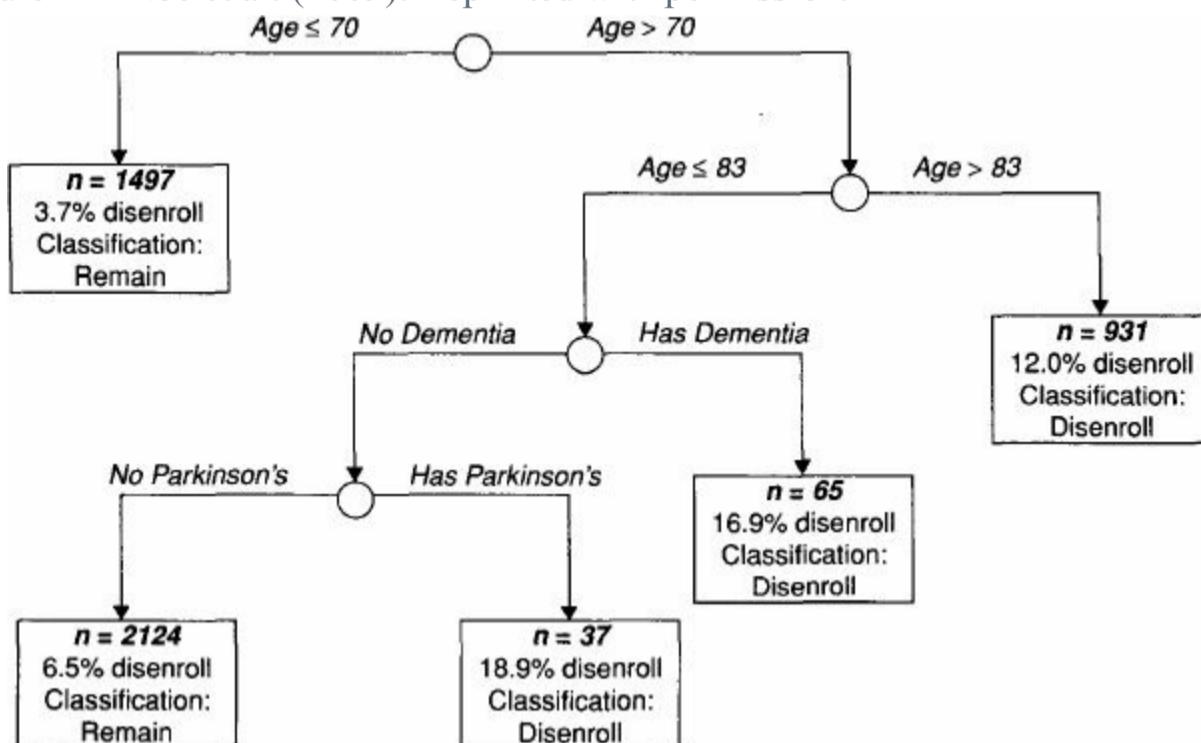
15.2.2 Example: Classification Tree for a Health Care Application

We use an example to illustrate the components of the classification tree method. Noe et al. (2009) predicted whether, over a one-year period, elderly subjects participating in an assisted-living program disenroll from the program to enter a nursing home. The sample consisted of 4654 individuals who had been enrolled in the program for at least a year and who did not die during that one-year period. Of this sample, 325 (7%) disenrolled from the program during the year.

[Figure 15.2](#) shows the classification tree. It summarizes responses to four questions with binary outcomes, listed next together with the counts having each response at a particular branch of the tree:

Figure 15.2 Classification tree for predicting disenrollment from an assisted-living program.

Source: Figure 2 in Noe et al. (2009). Reprinted with permission.



- Q1: Is the subject's age > 70? (3157 yes, 1497 no)
Q2: Is the subject's age > 83? (931 yes, 2226 no)
Q3: Does the subject have dementia? (65 yes, 2161 no)
Q4: Does the subject have Parkinson's disease? (37 yes, 2124 no)

As the tree indicates, those predicted to disenroll from the program were those of age >83, those of age >70 and \leq 83 having dementia, and those of age >70 and \leq 83 not having dementia but having Parkinson's disease. In summary, those of age >83 are predicted to disenroll, those of age \leq 70 are predicted to remain in the program, and those of age between 70 and 83 are predicted to disenroll if they have dementia or Parkinson's disease.

The points on the classification tree at which binary splits occur are called *nodes*. The initial node containing all the observations is the *root node*. The nodes beyond which no further splits occur, shown by boxes instead of circles in [Figure 15.2](#), are called *terminal nodes*. [Figure 15.2](#) has 9 nodes, of which 5 are terminal. The terminal nodes partition the entire sample into disjoint subsets.

15.2.3 How Does the Classification Tree Grow?

The method for constructing a binary classification tree uses a *recursive partitioning* algorithm for determining (1) how to choose the splitting variable at each node, (2) how to split a node on a chosen variable, and (3) how to declare a node to be terminal. Without going into detail, we now outline the main ideas. First, binary splits are used instead of multiway splits so the data do not get too fragmented too quickly. In any case, multiway splits can result from a series of binary splits, such as [Figure 15.2](#) does with age.

[Figure 15.2](#) predicts that $931 + 65 + 37 = 1033$ of the 4654 subjects disenroll. In reality, only 325 actually did disenroll. In terms of frequency of misclassification, the naive rule that predicts that *no one* would disenroll does better. The classification tree does not use this naive rule because the two types of misclassifications were treated differently in constructing the tree. Noe et al. (2009) focused on identifying subjects who would disenroll. They assigned a cost 13 times as high to predicting that someone would remain in the program who actually left it than to predicting that someone would leave the program who actually stayed. The relative misclassification cost for each possible prediction is a primary factor to determining the splits. Because of these differing costs, the method produces three terminal nodes identifying subjects as susceptible to disenroll, although the actual percentages of people who disenrolled in these nodes are all no greater than 18.9%. If instead the costs of the two types of misclassification were equal, at each terminal node the prediction would merely be the one with the smallest number of misclassifications.

The tree-structured classification method begins at the root node with all the sample subjects and first selects the best binary predictor of the response variable. In [Figure 15.2](#), age, which is continuous, is split into ≤ 70 and > 70 . This produces two new nodes, each of which are candidates for further binary splitting. For an ordinal variable or a quantitative variable such as age, the split takes the form of values falling above versus below a particular level. For a nominal variable, the split is based on ordering the categories by the sample proportions falling in the response category of interest and then using the same criterion to select a cutpoint to separate them into two sets of categories.

To find the first binary split, the algorithm forms a classification table of the form of [Table 15.2](#) for each possible binary split for each predictor variable. Ideally, two nodes would provide perfect prediction, with all observations in one row of the table falling in one column and all observations in the other row falling in the other column. The optimal split comes closest to this, in the sense of maximizing the difference between the deviance (based on a binomial likelihood function) for the model with a common probability for all observations and the model allowing two disjoint regions of x values each having a common probability. For a typical algorithm, this corresponds to using a statistical test, selecting the split that yields the smallest P -value in a test of the hypothesis that the created binary variable has no effect on the response. The significance can be judged after making some Bonferroni-type adjustment (Loh and Shih 1997, Sec. 7.5.2). The same procedure is then used with each new node.

Table 15.2 A Classification Table for a Predictor Split with the Disenrollment Response

Response Outcome		
Predictor x	$\hat{y} = 1$ (Disenroll)	$\hat{y} = 0$ (Remain)
Left node $x \leq c$	n_{11}	n_{12}
Right node $x > c$	n_{21}	n_{22}

The tree can continue growing until there are as many nodes as distinct sets of values of the predictors. In practice, this is overfitting, and a stopping rule is employed, such as stopping when any new node would have fewer than some fixed number of observations.

For a terminal node, the prediction taken is the response category that has the lowest misclassification cost. For example, consider the node of 931 subjects of age > 83 , of whom 112 disenrolled and 819 stayed. We treat the cost as 1 for misclassifying someone to disenroll who actually stays and 13 for misclassifying someone to stay who actually disenrolls. The

misclassification cost is then 819 if we predict that these 931 subjects disenroll and it is $13(112) = 1456$ if we predict that these 931 subjects stay. The misclassification cost is lower if we predict that they all disenroll, so this is the prediction for this terminal node.

15.2.4 Pruning a Tree and Checking Prediction Accuracy

For a classification tree to perform better for future prediction and not be overfitted to the data, some branches of the tree produced by the basic algorithm can be eliminated. This process is called *pruning*. One way to prune employs a measure of the quality of a tree that is an average of the quality of the terminal nodes, weighted by the proportion of observations at each such node. Let $p(t)$ denote the proportion of observations that occur at terminal node t . Let $c(t)$ denote the average misclassification cost at that node, which is the total misclassification cost for the predictions made at that terminal node divided by the number of subjects at that node. For example, for the 931 disenroll predictions at age >83 which account for the fraction $p(t) = 931/4654 = 0.20$ of the sample, we have $c(t) = 819/931 = 0.88$. Ideally we want a relatively simple tree that has good predictive accuracy. Thus, we could use a criterion corresponding to minimizing a measure such as

$$\sum_t p(t)c(t) + [\lambda \times (\text{number of terminal nodes})],$$

where the sum is taken over the terminal nodes and λ is a smoothing parameter.

The choice of λ reflects the bias/variance trade-off between fitting the data well (many terminal nodes, low bias) and having a parsimonious tree (relatively few terminal nodes, low variance). With $\lambda = 0$ we get the most complex possible tree. Generally, with very small λ , the data may be overfitted. As λ increases, more pruning occurs and the tree gets simpler. For $\lambda_1 < \lambda_2$, a tree with smoothing parameter λ_2 is nested within the tree with parameter λ_1 . Intervals of λ values result in the same tree. In practice, λ is chosen in an adaptive manner. Ideally, trees for different λ are tested on a separate validation sample to estimate their predictive accuracies. Several trees may have weighted misclassification cost for this validation sample that is near the minimum such cost. A tree is chosen that is relatively simple but has weighted misclassification cost close to the minimum. If a separate sample is not available, then a cross-validation method can use part of the original sample to suggest possible trees and the rest of the sample to test them.

Table 15.3 Classification Accuracy for Tree in Figure 15.2

		Observed Response Outcome	
		Disenroll	Remain
Classification Prediction			
Disenroll	130	903	
Remain	195	3426	

The classification accuracy is portrayed by a classification table. [Table 15.3](#) shows the table for this example. We can summarize such a table with sensitivity and specificity measures (Sections 2.1.3 and 6.3.3). The tree correctly predicts the proportion $130/(130 + 195) = 0.40$ of those who actually disenrolled and $3426/(903+3426) = 0.79$ of those who remained. An ROC curve can show how these rates vary as we vary the misclassification costs, thus affecting the predictions. The area under the ROC curve can be compared for various classification methods as a way of comparing their success rates (Hastie et al. 2009, Sec. 9.2). However, Hand (2009) argued that this measure is incoherent, because of different misclassification costs for different classification rules. He proposed an alternative approach based on averaged misclassification cost rather than averaged sensitivity.

For many data sets, some subjects are missing observations on at least one predictor variable. Various approaches can be used so those subjects enter the analysis. When considering a predictor for a particular split, a simple approach uses only observations for which that predictor is not missing. For a categorical predictor, instead adding a new category for missing can help reveal when missingness is not at random but is associated with a certain outcome (Hastie et al. 2009, p. 311).

15.2.5 Classification Trees Versus Logistic Regression

Classification trees provide a simple mechanism for using answers to a set of binary explanatory questions to predict a binary-response variable. A person can view the tree and clearly see which subjects have $y = 1$. Compared with logistic regression and other binary classification methods, tree-structured classification has the advantage of being easily understandable and useable by practitioners who have little understanding of basic statistics. Also, the trees do not require assumptions about the functional relationship between the response variable and the predictor variables. In particular, it is easier to detect potentially important interaction structure among the predictors, it is not necessary to prespecify categories for continuous predictors such as age, and the trees are invariant to monotone transformations of such predictors. The trees can more easily accommodate missing data on some predictors, and they rely on well-defined variable selection procedures, which is a thorny issue for logistic regression with a large number of explanatory variables.

A disadvantage of a classification tree compared with logistic regression modeling is the lack of smoothness caused by each terminal node of subjects being treated in the same way, since the region of explanatory variable values having $y = 1$ is a set of rectangular regions. In the above example, for instance, all subjects of age ≤ 70 are predicted to remain in the program, regardless of their values on other explanatory variables. If there truly is a simple linear structure for how the explanatory variables affect the response, as in a logistic regression model having only main effects, the tree will not help us discover this structure. Logistic regression has the advantage over classification trees and discriminant analysis of providing direct ways of summarizing effects of explanatory variables, through odds ratios. Moreover, those effects are all conditional on the other explanatory variables, whereas with the classification tree the displayed effects are mixed; the first split refers to a marginal effect, the second to a conditional effect given the first split, and so forth. Also, rather than relying on an automatic algorithm for forming a tree, in many applications it is better to use existing theory to suggest variables to use at particular levels of the hierarchy.

Finally, the classification tree method has low bias but high variance. There can be high variability in classification trees produced by different random samples from a common population, partly because of its hierarchical nature. Two samples that have different initial split may end up with very different trees because of the influence of the initial split on the way the tree evolves. Or, an optimal split early in the tree construction may cause the tree-constructing algorithm to miss another useful classifier.

Because of this variability and the segmentation into possibly very small groups that occurs with multiple splits, the classification tree method can require rather large sample sizes to work effectively. Even then, when the number of predictors is large, simpler methods such as nearest neighbor methods and linear discriminant methods that treat the explanatory variables as uncorrelated may have better classification performance. For example, see Dudoit et al. (2002), who compared various methods for classifying tumors using gene expression data. To reduce the high variability effect, L. Breiman proposed generalizations of classification trees. *Bagging* (a term that stands for “bootstrap aggregation”) is a method of averaging many trees, each constructed from an alternative sample that is generated from the original one using the bootstrap (Hastie et al. 2009, Sec. 8.7). *Random forest* ensembles of tree-classifiers also average trees, but select at each node a small group of input variables on which to consider splits, the goal being to reduce the correlation between trees (Hastie et al. 2009, Chap. 15). A disadvantage is that the overall contribution of a particular predictor is less clear than in ordinary classification trees or in logistic regression.

Because of the lack of smoothness, high variability, and atheoretic nature of classification trees, many researchers use this method mainly in an exploratory manner. Results of a classification tree analysis, combined with existing theory, can suggest logistic models to use in future research.

15.2.6 Support Vector Machines for Classification

In summary, Sections 15.1 and 15.2 suggest that (1) if it seems reasonable to assume normally distributed X with common covariance, simple linear discriminant analysis is appropriate for classification; (2) if X may be far from normal but logistic regression seems reasonable, we can use it for classification; (3) if X may interact in unknown ways to determine y but simple rectangular regions are desired for classification, then tree-structured methods are sensible.

Finally, a more complex method, *support vector machines*, has a decision boundary that can be highly irregular. For logistic regression, we've seen (Section 6.5) that perfect prediction and at least one infinite ML parameter estimate results when a hyperplane can separate the set of x values for which $y = 1$ from the set of x values for which $y = 0$. This hyperplane is a linear decision boundary. For classification purposes with future predictions, the optimal hyperplane has the maximum margin of separation between it and the nearest data points in the two sets of observations. In practice, we usually expect the sets to overlap, and such a linear decision boundary giving perfect predictions is not available. A support vector machine attempts to improve predictions over such hyperplane decision boundaries by producing nonlinear boundaries.

With support vector machines, the set of x values for which $y = 1$ is more complex than with linear discriminant analysis or with tree-structured classification. The set is determined by producing a linear boundary in a transformation of the space of explanatory variables, essentially replacing the predictors in a linear discriminant by some (possibly much larger) set of functions of them. The boundary depends on only a subset of the x_i values, which are the *support vectors* and fall on the margin hyperplanes. A kernel smoothing parameter controls the degree of nonlinearity, and a separate smoothing parameter controls the desired size of margin between the decision boundary and the nearest data points.

Hastie et al. (2009, p. 111) stated that ordinary linear discriminant analysis often performs well compared with more exotic methods. Even though linear discriminant analysis may have higher bias, it has the benefit of low variance because of its simplicity. Similarly, Hand (2006) argued that "simple methods typically yield performance almost as good as more sophisticated methods, to the extent that the difference in performance may be swamped by other sources of uncertainty." He noted, for example, that in practice the data points available for determining the classification rule are not randomly drawn from the same distribution to which the classifier will be applied, so statements about classifier accuracy need to take this into account. Also, interpretability is often an important requirement of a classification rule, and this favors simple methods.

15.3 CLUSTER ANALYSIS FOR CATEGORICAL DATA

The methods presented so far in this chapter have distinguished between response and explanatory variables. For example, discriminant analysis and classification trees are like logistic regression in using values on explanatory variables to classify observations into two well-defined groups that are the categories of the response variable. In some applications, such groups are not identified, but it is still relevant to sort observations into *clusters* of observations.

For example, in “market basket data” applications, a person’s observation is a vector of binary indicators in which a particular component indicates whether the person purchased the corresponding item. A consumer research study might want to identify groups of customers with similar buying behavior. Likewise, a company such as Google that provides Internet searching capability might seek to identify groups of people who have similar browsing behavior. A company such as Amazon recommends products to people based on a cluster *affinity analysis* that takes into account their purchase history and the history of other people who have bought the same items. A financial institution might try to detect outliers in purchases, such as in credit card fraud detection. In biology, clustering methods can organize plants or animals into groups according to observed features. With gene microarray data, clustering methods can identify clusters of genes with similar patterns of expression and may help to identify genes responsible for certain diseases (Hastie et al. 2009, Chap. 7; Dudoit et al. 2003). Clustering methods have even been used to group different brands of Scotch whisky (Lapointe and Legendre 1994).

15.3.1 Supervised Versus Unsupervised Learning

Discriminant analysis is sometimes referred to as *supervised learning*, because known classifications for some observations can be used (as if provided by a “supervisor”) to develop a discriminant function that can classify other observations measured on the same explanatory variables. By contrast, clustering methods are examples of *unsupervised learning*: The classification categories (the clusters) are unknown but features are observed that relate to the unobserved categories.

In this section we'll consider data sets consisting of n observations on a vector of p binary variables, the goal being to group those observations into a set of k clusters. Those clusters can be regarded as categories of an unknown variable. The number k itself may be unknown.

We can summarize the data as a 2^p contingency table that cross-classifies the n observations on the p binary variables. In some applications, such as market basket data, p may be very large. The table is then extremely sparse. It is more useful to express analyses directly in terms of the $n \times p$ data file of indicator variables, where row i shows the p binary responses (y_{i1}, \dots, y_{ip}) for observation i .

15.3.2 Measuring Dissimilarity Between Observations

Clustering methods use a measure of dissimilarity between observations. The clusters group together similar observations. Ideally, observations within a cluster have low dissimilarity whereas observations in different clusters have high dissimilarity. A clustering method is characterized by its dissimilarity measure and the algorithm for implementing the clustering.

For vectors of observations on p binary variables, [Table 15.4](#) summarizes the similarity and dissimilarity for observations h and i . There are a variables j in the vector for which $y_{hj} = y_{ij} = 1$, and d for which $y_{hj} = y_{ij} = 0$. A simple similarity measure for a pair of observations is the proportion of the $p = (a + b + c + d)$ variables that have a match, which is $(a + d)/(a + b + c + d)$. The corresponding dissimilarity measure subtracts the similarity measure from 1, giving the proportion $(b + c)/(a + b + c + d)$ for which the outcome differs.

Table 15.4 Cross Classification of Two Observations on p Binary Variables, where $p = (a + b + c + d)$

		Observation i
Observation h		
		1
1	a	b
0	c	d

In some applications, a common response of 1 is more relevant than a common response of 0. With market basket data, for example, each person's observation consists of a very high proportion of 0 entries (i.e., items *not* bought), so there is necessarily a high proportion of variables with a common outcome. Then, an asymmetric similarity measure may be more relevant. A popular similarity measure of this type is $a/(a + b + c)$, the number of variables coded as 1 for both observations divided by the number of variables that are coded as 1 for either or both observations. This is a special case of the *Jaccard index*, which for two sets is defined as the size of the intersection divided by the size of the union. The corresponding dissimilarity index is $(b + c)/(a + b + c)$.

Likewise, measures of similarity and dissimilarity can be defined for pairs of clusters of observations. The *average linkage* measures the average of the dissimilarities between all the pairs of observations, one from each cluster. Such measures do not account, however, for associations among the variables and treat them all identically. Alternatively, if the observations result from a probability sample, such as multinomial over the 2^p cells of the table cross-classifying the p variables, we could use a log-likelihood-based measure of dissimilarity. For example, the distance between two clusters could be defined as the decrease in the maximized log likelihood when we compare a model with separate parameter values for each cluster to a model with a common parameter value for the two clusters.

15.3.3 Clustering Algorithms: Partitions and Hierarchies

For a particular dissimilarity measure, two types of algorithms are commonly used to perform the clustering. One type *partitions* the observations in various ways and evaluates each partition according to some criterion. For a working partition at some stage for a k -cluster solution, the *medoid* of a cluster is the observation with smallest total dissimilarity to the other points in the cluster. The goal of *k -medoid clustering* is to seek an optimal partition in terms of minimizing the sum over the clusters of the total within-cluster dissimilarities between the observations and their medoids. For an initial partitioning, one algorithm assigns each observation to the cluster to which it has smallest dissimilarity with that cluster's medoid, then recomputes the medoids, and iterates. Kaufman and Rousseeuw (1990) proposed an alternative strategy that successively moves each medoid to an observation that is not currently one, then making the exchange that provides greatest reduction in the sum of the total within-cluster dissimilarities, continuing until no exchanges are found that provide an improvement.

The other main type of algorithm creates a *hierarchical* decomposition of the observations according to some criterion. The clusters at a particular level of the hierarchy result from merging clusters at the next level. At one extreme there is a single cluster of all observations and at the other extreme each observation forms its own cluster. The entire hierarchy portrays an ordered sequence of clusters. We can either create clusters by starting with each observation as its own cluster and merging them (*agglomerative clustering*) or instead start with all observations in a single cluster and at each stage divide an existing cluster into two clusters (*divisive clustering*). With agglomerative clustering, a step of the algorithm combines into a single cluster the pair of clusters having the smallest dissimilarity. With a hierarchical clustering method, a tree called a *dendrogram* displays the process of merging or dividing clusters. It portrays the grouping as a function of a metric such as the average dissimilarity between clusters being merged, and hence shows the clusters at each stage. The example in Section 15.3.4 illustrates.

Either clustering algorithm has advantages and disadvantages. For the agglomerative hierarchical approach, two samples from a population that combine clusters differently at an early stage may have quite different looking dendograms at a later stage. For the partitioning approach, it is usually computationally impractical to consider all possible partitions. It is necessary to either implement some stochastic element to the process or weaken the criterion used. Ultimate results may depend on the initial partition used, and it is sensible to try a few different ones (e.g., perhaps including the solution obtained with a hierarchical method) to increase the chance of finding the globally optimal solution. According to Hastie et al. (2009, p. 506), “specifying an appropriate dissimilarity measure is far more important in obtaining success with clustering than the choice of clustering algorithm.” For any algorithm, however, the clusters found may not reflect a true categorical classification but may merely be an artifact of that algorithm.

The number of clusters k is often unknown. Any algorithm, whether of a partitioning or a hierarchical nature, requires some termination condition for determining k . For example, agglomerative hierarchical clustering could keep combining clusters as long as the average dissimilarity between a pair of clusters to be combined is less than some particular fixed value. An informal way to choose k plots the value of the clustering criterion against k , looking for a natural break point where this changes substantially. Or, we could plot (against k) a summary such as the probability that the dissimilarity for a randomly selected within-cluster pair of observations is smaller than the dissimilarity for a randomly selected between-cluster pair. An adaptation of the Goodman and Kruskal gamma measure takes the difference between this *concordance probability* and a *discordance probability*, divided by their sum (Baker and Hubert 1975).

Partitioning and hierarchical clustering methods need not assume a probability model for the data. Some clustering methods, though, are probabilistic model-based. Such approaches usually assume that the observations come from a k -component mixture distribution of some type. An example is the

latent class model of Section 14.1. With it, each observation is not actually assigned a cluster but rather a probability distribution over the clusters (latent classes) to which it could belong (Fraley and Raftery 2002, Magidson and Vermunt 2004). In practice, observations are typically assigned to the cluster for which the probability value is highest. This model-based approach applies also directly to multcategory response data.

With very high-dimensional data, such as market basket data or DNA microarray data, the challenges to clustering are many, regardless of the type of data. There may be many irrelevant variables that have the impact of masking clusters, as clusters might exist only for a very small subset of the variables. Dissimilarity measures then are less meaningful, as observations may be close for the most relevant variables but the curse of dimensionality may put them far apart in high-dimensional space. To attempt to account for this, clustering can be attempted in various subspaces by clustering observations on *subsets* of variables rather than all of them simultaneously (Brusco 2004, Friedman and Meulman 2004).

15.3.4 Example: Clustering States on Election Results

The text website has a 51×8 matrix, with a row for each U.S. state and D.C., that shows the party (Democrat or Republican) that won the electoral votes for that state for each presidential election between 1980 and 2008. [Table 15.5](#) shows an excerpt from that table.

Table 15.5 Statewide Data on Party (Dem = Democrat, Rep = Republican) Winning Electoral Votes in Presidential Elections between 1980 and 2008

State	1980	1984	1988	1992	1996	2000	2004	2008
Arizona	Rep	Rep	Rep	Rep	Dem	Rep	Rep	Rep
California	Rep	Rep	Rep	Dem	Dem	Dem	Dem	Dem
Colorado	Rep	Rep	Rep	Dem	Rep	Rep	Rep	Dem
Florida	Rep	Rep	Rep	Rep	Dem	Rep	Rep	Dem
Illinois	Rep	Rep	Rep	Dem	Dem	Dem	Dem	Dem
Massachusetts	Rep	Rep	Dem	Dem	Dem	Dem	Dem	Dem
Minnesota	Dem	Rep	Dem	Dem	Dem	Dem	Dem	Dem
Missouri	Rep	Rep	Rep	Dem	Dem	Rep	Rep	Rep
New Mexico	Rep	Rep	Rep	Dem	Dem	Dem	Rep	Dem
New York	Rep	Rep	Dem	Dem	Dem	Dem	Dem	Dem
Ohio	Rep	Rep	Rep	Dem	Dem	Rep	Rep	Dem
Texas	Rep							
Virginia	Rep	Dem						
Wyoming	Rep							

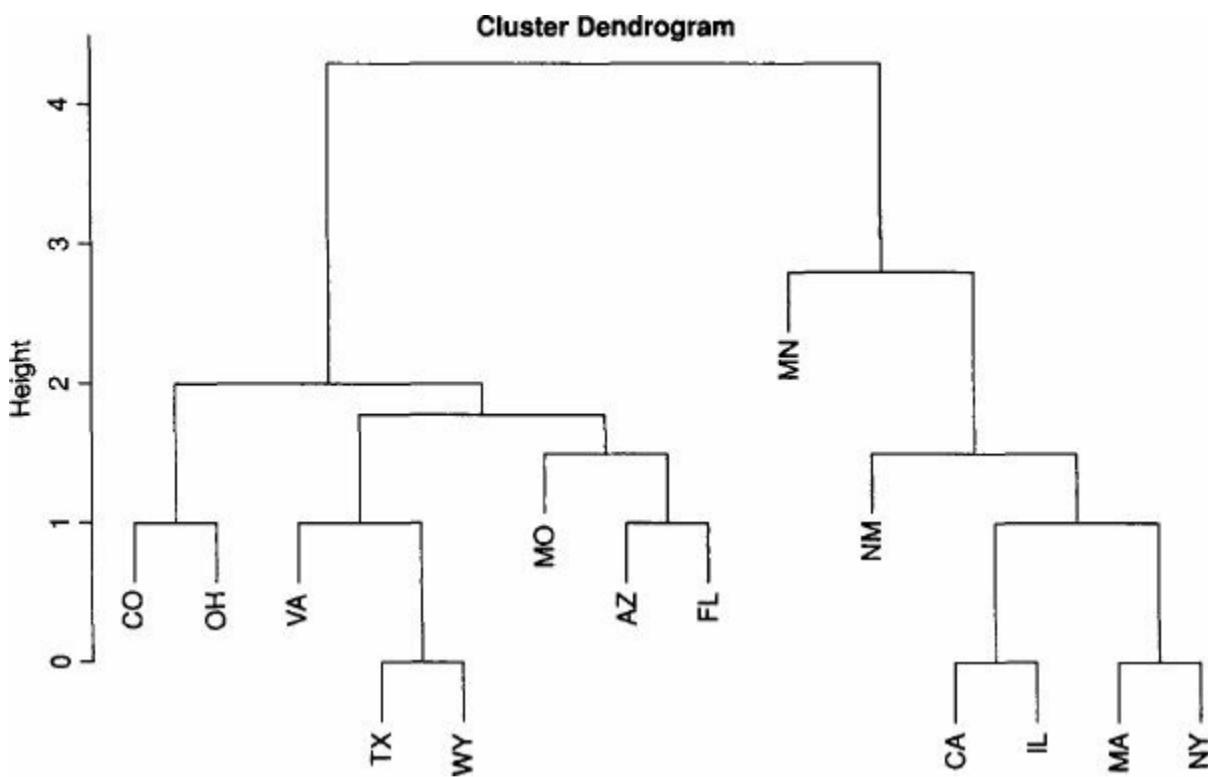
Source: Complete data at www.stat.ufl.edu/~aa/cda/cda.html.

We measure dissimilarity using the number of elections on which the states differ. States with identical vectors of responses, such as Massachusetts and New York, have dissimilarity values of 0. By contrast, Minnesota and Texas differ in the outcome for every election, so the dissimilarity is 8. Minnesota and Virginia agree in only 1 of the 8 elections, so their dissimilarity is 7.

The agglomerative hierarchical algorithm starts with 51 clusters, one for each state and D.C. At the first step, states are combined that have the minimum dissimilarity, which in this case consists of states such as Massachusetts and New York that have dissimilarity of 0. At that step, eight clusters of states have dissimilarity of 0 for all pairs within each cluster. At the next step, clusters are combined with the next smallest dissimilarity, such as those two states with California, for which the dissimilarity is 1. By the stage at which there are only two clusters, one cluster has 21 states and D.C. that tend to vote Democrat, and the other cluster has 29 states that tend to vote Republican.

To more easily portray the agglomerative cluster-forming process as well as a dendrogram for displaying results, we redo the analysis using only the data shown in [Table 15.5](#) for 14 states. [Figure 15.3](#) shows the dendrogram. The bottom nodes of the figure show the initial 14 clusters. The vertical scale shows the average dissimilarity between clusters being joined. At the first stage, three pairs of clusters are joined, of two states each, that have dissimilarities of 0. At the next stage, the (California, Illinois) cluster is joined with the (Massachusetts, New York) cluster, the average dissimilarity between those two clusters being 1. At the same stage, Virginia is joined with the (Texas, Wyoming) cluster, Colorado and Ohio are joined in a cluster, and Arizona and Florida are joined in a cluster. At this stage there are seven clusters, these four plus the single-state clusters of Missouri, New Mexico, and Minnesota.

[Figure 15.3](#) Dendrogram (produced using *dist* and *hclust* functions in R) for cluster analysis of 14 states according to presidential election results, for data in [Table 15.5](#).



The top of the dendrogram is the joining of all states into a single cluster. The two-cluster solution, below it, shows the Republican-leaning cluster (Colorado, Ohio, Virginia, Texas, Wyoming, Missouri, Arizona, Florida) and the Democrat-leaning cluster (Minnesota, New Mexico, California, Illinois, Massachusetts, New York). At the two-cluster step, when Minnesota is joined with five other states, the average dissimilarity between Minnesota and the other five states is $14/5 = 2.8$.

NOTES

Section 15.1: Classification: Linear Discriminant Analysis

15.1 Discriminant versus logistic: For more on discriminant analysis, see Hastie et al. (2009, Chap. 4), Lin et al. (2010), McLachlan (2004), and Tutz (2011, Chap. 15). For classification when normality does not hold, Anderson (1975), Bull and Donner (1987), Efron (1975), McLachlan (2004, Sec. 8.5), and Press and Wilson (1978) compared methods, generally rating logistic regression more favorably than discriminant analysis.

15.2 Discriminant generalizations: The prediction rule in discriminant analysis can be amended to take into account different misclassification costs for the two types of misclassifications or to minimize a risk function based on a penalty function. See Eguchi and Copas (2002). It generalizes also to handle multiple classes, regularized covariance matrices and lasso-type penalties, assuming a mixture of normal distributions for each category, penalizing the coefficients to make them smoother, and nonparametric estimation of the distribution of ($X \setminus Y = j$) or of the form of the regression. See Hastie et al. (2009, Sec. 12.4) and Witten and Tibshirani (2011). Articles dealing with classification methods for large p include Bickel and Levina (2004), Fan and Fan (2008), Friedman (1989), Mai et al. (2012 and references therein), Tibshirani et al. (2003), and Wu et al. (2009). Fan and Fan (2008) showed a way to quantify the impact of dimensionality on classification.

Section 15.2: Classification: Tree-Structured Prediction

15.3 Trees/extensions: For more about tree-structured classification and its generalizations, see Breiman et al. (1984), Hastie et al. (2009, Sec. 9.2), Loh (2002), Loh and Shih (1997), Tutz (2011, Chap. 11), and Zhang and Singer (2010). See Zhang (1998) for extensions to multiple binary responses, Piccarreta (2008) for ordinal responses, and Meulman (2003) for an interesting overview. Hastie et al. (2009) is a good but technical reference for various “machine learning” methods. See Chapter 12 for support vector machines. See Azzalini and Scarpa (2012) for a less technical introduction to data mining methods. Blanchard et al. (2008) studied the support vector machines algorithm from a statistical perspective. Li (2010) proposed a boosting algorithm for multiclass tree-based classification.

Section 15.3: Cluster Analysis for Categorical Data

15.4 Clustering extensions: For examples and details about clustering algorithms, although mainly for continuous variables, see Azzalini and Scarpa (2012), Everitt et al. (2011), Fraley and Raftery (2002), Hastie et al. (2009, Sec. 14.3), and Kaufman and Rousseeuw (1990). Booth et al. (2008) proposed a multilevel linear mixed model, having the feature that observations from the same cluster are correlated because they share cluster-specific random effects. One of the parameters in the model is the true underlying partition of the data, and the posterior distribution of this parameter is used to cluster the data. Hitchcock and Chen (2008) showed advantages to smoothing the dissimilarities before clustering binary data, by smoothing the proportion estimates of the agreements and disagreements on the p variables. Friedman and Meulman (2004) and Hunt and Jorgensen (1999) considered clustering with mixed categorical and continuous variables.

EXERCISES

Applications

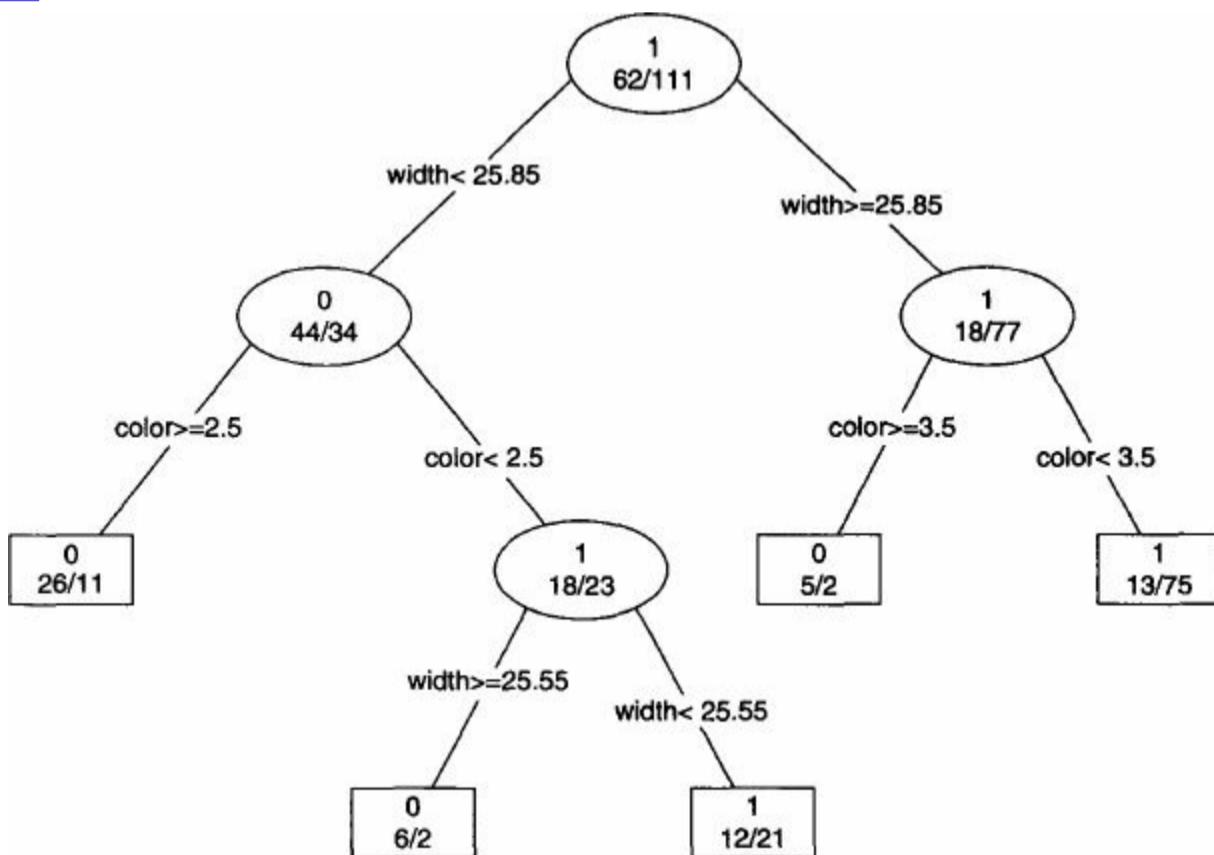
15.1 Refer to the classic use of discriminant analysis by Fisher (1936) for Iris flower data as discussed in the article “Iris flower data set” at Wikipedia. Conduct a linear discriminant analysis using the data given there for the versicolor and virginica species, with sepal length and petal length as explanatory variables. Use $\pi_0 = 0.50$. Report the linear discriminant function and show the cross-validated classification table.

15.2 For the classification tree shown in [Figure 15.2](#), explain what the prediction would be at each terminal node if the cost of misclassifying a person as remaining in the program when he/she actually disenroll were (a) 7 times and (b) equal to, the cost of misclassifying a person as disenrolling when they actually remain.

15.3 Refer to the previous exercise. Suppose you were to conduct a binary regression analysis of these data. Summarize the advantages and disadvantages of this approach compared with the classification tree analysis.

15.4 [Figure 15.4](#) is a classification tree obtained for the horseshoe crab data with explanatory variables width and quantitative color (as in Section 15.1.2), using the *rpart* and *prune* functions in R, with the complexity parameter set at 0.02 for the pruning. For example, of the 88 crabs with width ≥ 25.85 cm and color in the three lowest categories (1, 2, 3), 88 crabs were predicted to have satellites; in fact, 75 had them but 13 did not.

[Figure 15.4](#) Pruned classification tree for horseshoe crabs.



- Summarize what the terminal nodes tell you.
- Construct the classification table. (This is not strictly comparable to [Table 15.1](#) for logistic modeling and discriminant analysis, because that table used cross-validation.)
- For crabs having width < 25.85 cm, explain how the two terminal nodes having color in one of the two lightest categories give predictions (relative to each other) that contradict what the logistic model fit suggests.

15.5 Use the classification tree method with the horseshoe crab data (available at the text website), assuming equal misclassification costs and all four explanatory variables. Specify the criteria you chose to build and prune the tree. Explain how to interpret the pruned tree, and explain how (if at all) it conflicts with the results from the model-building of Section 6.1.4.

15.6 For the previous exercise, compare this method's classification accuracy for these data to that for (a) a logistic regression model with the same predictors and (b) a linear discriminant analysis with the same predictors. To keep results comparable, either use cross-validation in all cases or do not use it in all cases.

15.7 Using the *spam* data set at the website www-stat.stanford.edu/~tibs/ElemStatLearn for Hastie et al. (2009), use two methods presented in this chapter to classify whether a given email is spam. Explain how you implemented the methods, form classification tables, and summarize results.

15.8 For the cluster analysis example in Section 15.3.4, the observations were the same for New Jersey and Pennsylvania as Illinois, and the same for North Carolina as Virginia. Conduct and interpret a cluster analysis using these three states together with the 14 states in [Table 15.5](#).

15.9 For the previous exercise, conduct a cluster analysis using the full data set at the text website. Explain how you implemented the method, and interpret results.

15.10 The grounds on which a divorce of a marriage can be sought vary from state to state. [Table 15.6](#) shows data for eight states. The complete data are at the text website.

Table 15.6 Statewide Data on Grounds for Divorce, Where 1 = Yes and 0 = No

State	Grounds for Divorce								
	1	2	3	4	5	6	7	8	9
California	1	0	0	0	0	0	0	1	0
Florida	1	0	0	0	0	0	0	1	0
Illinois	0	1	1	0	1	1	1	0	0
Massachusetts	1	1	1	1	1	1	1	0	1
Michigan	1	0	0	0	0	0	0	0	0
New York	0	1	1	0	0	1	0	0	1
Texas	1	1	1	0	0	1	0	1	1
Washington	1	0	0	0	0	0	0	0	1

Note: The grounds are (1) incompatibility, (2) mental cruelty, (3) desertion, (4) nonsupport, (5) alcohol abuse, (6) felony, (7) impotence, (8) insanity, (9) separation.

Source: From p. 1516 of *SAT/STAT 9.2 User's Guide: The DISTANCE Procedure*, © 2008 SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc.

- a. For a cluster analysis to identify groups of similar states, does it make sense to use a symmetric dissimilarity index, such as the proportion of grounds that differ, or an asymmetric measure such as the Jaccard dissimilarity index? Explain.
- b. For the dissimilarity method you chose, show the first two steps of an agglomerative hierarchical approach for the observations in [Table 15.6](#).

15.11 For the previous exercise, conduct a cluster analysis using the full data set at the text website.

- a. Using the Jaccard dissimilarity index, show that the nine-cluster solution has four clusters with single states, a cluster of ten states that have the same responses as Michigan, a cluster of four states that have the same responses as Florida and California, a cluster of four states that have the same responses as Washington, a cluster of 25 states including Massachusetts and Texas, and a cluster of three states including New York.
- b. Show the result of a two-cluster solution. Explain your choices for implementing the method, and display the dendrogram and interpret results.

15.12 Project: Go to a site with large data files, such as the UCI Machine Learning Repository (archive.ics.uci.edu/ml) or Yahoo! Webscope (webscope.sandbox.yahoo.com). Find a data set of interest to you that has a categorical response variable. Use at least one method presented in this chapter to analyze the data. Summarize your analyses in a two-page report, attaching an appendix showing your use of software.

Theory and Methods

15.13 Assuming that $(X|Y=j)$ has a multivariate $N(\boldsymbol{\mu}_j, \Sigma)$ distribution, $j = 0, 1$, derive the logistic

expression for $[P(Y=1|x)]$ given at the beginning of Section 15.1.1.

15.14 For a binary classification tree, explain why the number of nodes T relates to the number of terminal nodes \tilde{T} by $T = 2\tilde{T} - 1$.

15.15 For applications of cluster analysis such as to market basket data with extremely large p and a very high proportion of 0 responses, explain why the dissimilarity index $(b + c)/p$ (in the notation of [Table 15.4](#)) may not be appropriate.

15.16 For a 2×2 table for two binary variables, what clusters would be formed in a cluster analysis, under the constraint that two observations in the same cluster must have a dissimilarity value no greater than 0?

15.17 Explain the similarities and differences between cluster analysis and latent class analysis, in terms of sampling assumptions for the methods and in terms of the sorts of conclusions that are reached. Illustrate using the full data at the text website for the election results example in Section 15.3.4.

15.18 Do a literature search and write a two-page paper describing cluster analysis methods that are available when the observed features are multicategory rather than binary. For any method described, explain whether it treats variables as nominal or ordinal.

15.19 Based on reading appropriate literature, prepare a two-page report summarizing the **(a)** bagging or **(b)** random forest approach to classification trees. In the final paragraph of your report specify the method's advantages and disadvantages compared with other methods.

CHAPTER 16

Large- and Small-Sample Theory for Multinomial Models

This chapter gives a unified presentation of the large-sample theory and small-sample theory that we've used in this book for parametric models for categorical data. The primary emphasis is on multinomial models for contingency tables.

In Section 16.1 we review and extend the *delta method* for deriving large-sample normal distributions for many statistics. In Section 16.2 we apply the delta method to estimators of parameters in models for contingency tables, later illustrated in Section 16.4 for logistic and loglinear models. In Section 16.3 we derive large-sample distributions of cell residuals and the X^2 and G^2 goodness-of-fit statistics. We'll see that powerful results can follow from simple mathematical ideas, such as Taylor series expansions. In Sections 16.5 and 16.6 we present the theory for small-sample tests and confidence intervals for proportions and parameters for contingency tables. The emphasis throughout is on ML inference, but the final section mentions alternative approaches that have similar large-sample properties.

The results in this chapter have a long history. Pearson (1900) derived the limiting chi-squared distribution of X^2 for testing a specified multinomial distribution. Fisher (1922, 1924) showed the degrees of freedom adjustment when multinomial probabilities are functions of unknown parameters. Cramér (1946, pp. 424–434) formally proved this result, under the assumption that ML estimators of the parameters are consistent. Rao (1957) proved consistency of the ML estimators and derived their asymptotic distribution under general conditions. Birch (1964a) proved these results under weaker conditions. For small samples, significance tests generalize Fisher's (1935a) conditional approach for Fisher's exact test, and confidence intervals generalize work by Clopper and Pearson (1934) using small-sample distributions such as the binomial.

16.1 DELTA METHOD

Suppose that a statistic used to estimate a parameter has a large-sample normal distribution. In this section we show that many functions of that statistic are also asymptotically normal.

16.1.1 O , o Rates of Convergence

Big O and *little o* notation is useful for describing limiting behavior of sequences. For real numbers $\{z_n\}$, the little o notation $o(z_n)$ represents a term that has *smaller* order than z_n as $n \rightarrow \infty$, in the sense that $o(z_n)/z_n \rightarrow 0$ as $n \rightarrow \infty$. For instance, \sqrt{n} is $o(n)$ as $n \rightarrow \infty$, since $\sqrt{n}/n \rightarrow 0$ as $n \rightarrow \infty$. A sequence that is $o(1)$ satisfies $o(1)/1 = o(1) \rightarrow 0$; for instance, $n^{-1/2}$ is $o(1)$ as $n \rightarrow \infty$.

The big O notation $O(z_n)$ represents terms that have the *same* order of magnitude as z_n , in the sense that $|O(z_n)/z_n|$ is bounded as $n \rightarrow \infty$. For instance, $(3/n) + (8/n^2)$ is $O(n^{-1})$ as $n \rightarrow \infty$; dividing it by n^{-1} gives a ratio that takes value close to 3 for large n .

Similar notation applies to sequences of random variables. This notation uses a subscript p to indicate that the sequence has probabilistic rather than deterministic behavior. The symbol $o_p(z_n)$ denotes a random variable of *smaller* order than z_n for large n , in the sense that $o_p(z_n)/z_n$ converges in probability to 0; that is, for any fixed $\epsilon > 0$, $P(|o_p(z_n)/z_n| \leq \epsilon) \rightarrow 1$ as $n \rightarrow \infty$. The notation $O_p(z_n)$ represents a random variable such that for every $\epsilon > 0$, there is a constant K and an integer n_0 such that $P[|O_p(z_n)/z_n| < K] > 1 - \epsilon$ for all $n > n_0$.

For the sample mean \bar{y}_n of n independent observations Y_1, \dots, Y_n from a distribution having $E(Y_i) = \mu$, $(\bar{y}_n - \mu) = o_p(1)$, since $(\bar{y}_n - \mu)/1$ converges in probability to 0 as $n \rightarrow \infty$ by the law of large numbers. By Tchebychev's inequality, the difference between a random variable and its expected value has the same order of magnitude as the standard deviation of that random variable. Since $\bar{y}_n - \mu$ has standard deviation σ/\sqrt{n} , $(\bar{y}_n - \mu) = O_p(n^{-1/2})$.

A random variable that is $O_p(n^{-1/2})$ is also $o_p(1)$. An example is $(\bar{y}_n - \mu)$. Multiplication affects the order in a natural manner (Exercise 16.5). If the difference between two random variables is $o_p(1)$ as $n \rightarrow \infty$, Slutsky's theorem states that those random variables have the same limiting distribution.

16.1.2 Delta Method for a Function of a Random Variable

Let T_n denote a statistic, the subscript expressing its dependence on the sample size n . For large samples, suppose T_n has approximately a normal distribution with mean θ and standard error σ/\sqrt{n} . More precisely, as $n \rightarrow \infty$, the cdf of $\sqrt{n}(T_n - \theta)$ converges to a $N(0, \sigma^2)$ cdf. This limiting behavior is an example of *convergence in distribution*, denoted by

$$(16.1) \quad \sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2).$$

For a function g , we now derive the limiting distribution of $g(T_n)$. Suppose that g is at least twice differentiable at θ . By the Taylor series expansion of $g(t)$ in a neighborhood of θ , for some θ^* between t and θ ,

$$\begin{aligned} g(t) &= g(\theta) + (t - \theta)g'(\theta) + (t - \theta)^2 g''(\theta^*)/2 \\ &= g(\theta) + (t - \theta)g'(\theta) + O(|t - \theta|^2). \end{aligned}$$

Substituting the random variable T_n for t ,

$$\begin{aligned} \sqrt{n}[g(T_n) - g(\theta)] &= \sqrt{n}(T_n - \theta)g'(\theta) + \sqrt{n}O(|T_n - \theta|^2) \\ (16.2) \quad &= \sqrt{n}(T_n - \theta)g'(\theta) + O_p(n^{-1/2}) \end{aligned}$$

since

$$\sqrt{n}O(|T_n - \theta|^2) = \sqrt{n}O(O_p(n^{-1})) = O_p(n^{-1/2}).$$

The $O_p(n^{-1/2})$ term is asymptotically negligible, so $\sqrt{n}[g(T_n) - g(\theta)]$ has the same limiting distribution as $\sqrt{n}(T_n - \theta)g'(\theta)$; that is, $g(T_n) - g(\theta)$ behaves like the constant multiple $g'(\theta)$ of $(T_n - \theta)$. Now, $(T_n - \theta)$ is approximately normal with variance σ^2/n . Thus, $g(T_n) - g(\theta)$ is approximately normal with variance $\sigma^2[g'(\theta)]^2/n$. More precisely,

$$(16.3) \quad \sqrt{n}[g(T_n) - g(\theta)] \xrightarrow{d} N(0, \sigma^2[g'(\theta)]^2).$$

[Figure 3.1](#) illustrated this result, and in Section 3.1.6 we applied it to the sample logit.

Result (16.3) is called the *delta method* for obtaining asymptotic distributions. Since $\sigma^2 = \sigma^2(\theta)$ and $g'(\theta)$ usually depends on θ , the asymptotic variance is unknown. Let $\sigma^2(T_n)$ and $g'(T_n)$ denote these terms evaluated at the sample estimator T_n of θ . When $g'(\cdot)$ and $\sigma = \sigma(\cdot)$ are continuous at θ , then $\sigma(T_n)g'(T_n)$ is a consistent estimator of $\sigma(\theta)g'(\theta)$. Thus, Wald confidence intervals and tests use the result that $\sqrt{n}[g(T_n) - g(\theta)]/\sigma(T_n)|g'(T_n)|$ is asymptotically standard normal. For instance,

$$g(T_n) \pm z_{\alpha/2} \sigma(T_n)|g'(T_n)|/\sqrt{n}$$

is a large-sample $100(1 - \alpha)\%$ Wald confidence interval for $g(\theta)$.

16.1.3 Delta Method for a Function of a Random Vector

The delta method generalizes to functions of random *vectors*. Suppose that $\mathbf{T}_n = (T_{n1}, \dots, T_{nN})^T$ is asymptotically multivariate normal with mean $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N)^T$ and covariance matrix Σ/n . Suppose that $g(t_1, \dots, t_N)$ has a nonzero differential $\boldsymbol{\phi} = (\phi_1, \dots, \phi_N)^T$ at $\boldsymbol{\theta}$, where

$$\phi_i = \left. \frac{\partial g}{\partial t_i} \right|_{t=\theta}.$$

Then,

$$(16.4) \quad \sqrt{n}[g(\mathbf{T}_n) - g(\boldsymbol{\theta})] \xrightarrow{d} N(0, \boldsymbol{\phi}^T \boldsymbol{\Sigma} \boldsymbol{\phi}).$$

For large n , $g(\mathbf{T}_n)$ has distribution similar to the normal with mean $g(\boldsymbol{\theta})$ and variance $\boldsymbol{\pi}^T \boldsymbol{\Sigma} \boldsymbol{\pi}/n$.

The proof of (16.4) follows from the expansion

$$g(\mathbf{T}_n) - g(\boldsymbol{\theta}) = (\mathbf{T}_n - \boldsymbol{\theta})^T \boldsymbol{\phi} + o(\|\mathbf{T}_n - \boldsymbol{\theta}\|),$$

where $\|z\| = \sqrt{\sum_i z_i^2}$ denotes the length of vector z . For large n , $g(\mathbf{T}_n) - g(\boldsymbol{\theta})$ behaves like a linear function of the approximately normal random vector $(\mathbf{T}_n - \boldsymbol{\theta})$. Thus, it itself is approximately normal.

16.1.4 Asymptotic Normality of Functions of Multinomial Counts

The delta method for random vectors implies asymptotic normality of many functions of multinomial cell counts (n_1, \dots, n_N) in contingency tables with cell probabilities $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)^T$. Let $\mathbf{p} = (p_1, \dots, p_N)^T$ denote the sample proportions, where $p_i = n_i/n$ with $n = n_1 + \dots + n_N$. Denote observation i by $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iN})$, where $Y_{ij} = 1$ if it falls in cell j and $Y_{ij} = 0$ otherwise, $i = 1, \dots, n$. Since each observation falls in only one cell, $\sum_j Y_{ij} = 1$ and $Y_{ij} Y_{ik} = 0$ when $j \neq k$. Also, $p_j = \sum_i Y_{ij}/n$, and

$$E(Y_{ij}) = P(Y_{ij} = 1) = \pi_j = E(Y_{ij}^2), \quad E(Y_{ij} Y_{ik}) = 0 \quad \text{if } j \neq k.$$

It follows that

$$E(\mathbf{Y}_i) = \boldsymbol{\pi} \quad \text{and} \quad \text{cov}(\mathbf{Y}_i) = \boldsymbol{\Sigma}, \quad i = 1, \dots, n,$$

where $\boldsymbol{\Sigma} = (\sigma_{jk})$ with

$$\sigma_{jj} = \text{var}(Y_{ij}) = E(Y_{ij}^2) - [E(Y_{ij})]^2 = \pi_j(1 - \pi_j),$$

$$\sigma_{jk} = \text{cov}(Y_{ij}, Y_{ik}) = E(Y_{ij} Y_{ik}) - E(Y_{ij})E(Y_{ik}) = -\pi_j \pi_k \quad \text{for } j \neq k.$$

The matrix $\boldsymbol{\Sigma}$ has form

$$\boldsymbol{\Sigma} = \text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi} \boldsymbol{\pi}^T,$$

where $\text{Diag}(\boldsymbol{\pi})$ is the diagonal matrix with the elements of $\boldsymbol{\pi}$ on the main diagonal.

Since $\mathbf{p} = (\sum_i Y_i)/n$ is a sample mean of n independent observations,

$$(16.5) \quad \text{cov}(\mathbf{p}) = [\text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi} \boldsymbol{\pi}^T]/n.$$

This covariance matrix is singular, because of the linear dependence $\sum_i p_i = 1$. The multivariate central limit theorem (Rao 1973, p. 128) implies

$$(16.6) \quad \sqrt{n}(\mathbf{p} - \boldsymbol{\pi}) \xrightarrow{d} N[\mathbf{0}, \text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi} \boldsymbol{\pi}^T].$$

By the delta method, functions of \mathbf{p} having nonzero differential at $\boldsymbol{\pi}$ are also asymptotically normal. Let $g(t_1, \dots, t_N)$ be a differentiable function, and let $\phi_i = \partial g / \partial \pi_i$ denote $\partial g / \partial t_i$ evaluated at $t = \boldsymbol{\pi}$. By the delta method (16.4),

$$(16.7) \quad \sqrt{n}[g(\mathbf{p}) - g(\boldsymbol{\pi})] \xrightarrow{d} N(0, \boldsymbol{\phi}^T [\text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi} \boldsymbol{\pi}^T] \boldsymbol{\phi}).$$

The asymptotic variance equals

$$\boldsymbol{\phi}^T \text{Diag}(\boldsymbol{\pi}) \boldsymbol{\phi} - (\boldsymbol{\phi}^T \boldsymbol{\pi})^2 = \sum_i \pi_i \phi_i^2 - \left(\sum_i \pi_i \phi_i \right)^2.$$

In Section 3.1.7 we used this formula to derive the large-sample variance of the sample log odds ratio.

16.1.5 Delta Method for a Vector Function of a Random Vector

The delta method generalizes further to a *vector* of functions of an asymptotically normal random vector. Let $\mathbf{g}(\mathbf{t}) = (g_1(\mathbf{t}), \dots, g_q(\mathbf{t}))^T$ and let $(\partial \mathbf{g}/\partial \boldsymbol{\theta})$ denote the $q \times N$ Jacobian matrix for which the entry in row i and column j is $\partial g_i(\mathbf{t})/\partial t_j$ evaluated at $\mathbf{t} = \boldsymbol{\theta}$. Then,

$$(16.8) \quad \sqrt{n} [\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\theta})] \xrightarrow{d} N[\mathbf{0}, (\partial \mathbf{g}/\partial \boldsymbol{\theta}) \boldsymbol{\Sigma} (\partial \mathbf{g}/\partial \boldsymbol{\theta})^T].$$

The rank of the limiting normal distribution equals the rank of the asymptotic covariance matrix.

This expression is useful for finding large-sample joint distributions. For instance, from (16.6), (16.7), and (16.8), the asymptotic joint distribution of several functions of multinomial proportions has covariance matrix of the form

$$\text{asympt. cov}\{\sqrt{n} [\mathbf{g}(\mathbf{p}) - \mathbf{g}(\boldsymbol{\pi})]\} = \boldsymbol{\Phi} [\text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi} \boldsymbol{\pi}^T] \boldsymbol{\Phi}^T,$$

where $\boldsymbol{\Phi}$ is the Jacobian $(\partial \mathbf{g}/\partial \boldsymbol{\pi})$.

16.1.6 Joint Asymptotic Normality of Log Odds Ratios

We illustrate formula (16.8) by finding the asymptotic joint distribution of a set of log odds ratios in a contingency table. Let $\mathbf{g}(\boldsymbol{\pi}) = \log(\boldsymbol{\pi})$ denote the vector of natural logs of cell probabilities, for which

$$\partial \mathbf{g} / \partial \boldsymbol{\pi} = \text{Diag}(\boldsymbol{\pi})^{-1}.$$

The covariance of the asymptotic distribution of $\sqrt{n}[\log(\mathbf{p}) - \log(\boldsymbol{\pi})]$ is

$$\text{Diag}(\boldsymbol{\pi})^{-1} [\text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi} \boldsymbol{\pi}^T] \text{Diag}(\boldsymbol{\pi})^{-1} = \text{Diag}(\boldsymbol{\pi})^{-1} - \mathbf{1}\mathbf{1}^T,$$

where $\mathbf{1}$ is an $N \times 1$ vector of 1 elements.

For a $q \times N$ matrix of constants \mathbf{C} , it follows that

$$(16.9) \quad \sqrt{n} \mathbf{C} [\log(\mathbf{p}) - \log(\boldsymbol{\pi})] \xrightarrow{d} N[\mathbf{0}, \mathbf{C} \text{Diag}(\boldsymbol{\pi})^{-1} \mathbf{C}^T - \mathbf{C} \mathbf{1} \mathbf{1}^T \mathbf{C}^T].$$

Now, suppose $\mathbf{C} \log(\mathbf{p})$ is a set of sample log odds ratios. Then, each row of \mathbf{C} contains zeros except for two +1 elements and two -1 elements in the positions multiplied by the relevant elements of $\log(\mathbf{p})$ to form the given log odds ratio. The second term in the covariance matrix in (16.9) is then zero. If a particular odds ratio uses the cells numbered h , i , j , and k , the variance of the asymptotic distribution is

$$\text{asympt. var}[\sqrt{n} (\text{sample log odds ratio})] = \pi_h^{-1} + \pi_i^{-1} + \pi_j^{-1} + \pi_k^{-1}.$$

When two log odds ratios have no cells in common, their asymptotic covariance in the limiting normal distribution equals zero.

16.2 ASYMPTOTIC DISTRIBUTIONS OF ESTIMATORS OF MODEL PARAMETERS AND CELL PROBABILITIES

We now derive basic results of large-sample model-based inference for contingency tables. The delta method is the key tool. The derivations apply to a single multinomial distribution, but extend directly to products of multinomials for independent samples.

The observations are counts $\mathbf{n} = (n_1, \dots, n_N)^T$ in N cells of a contingency table. The asymptotics regard N as fixed and let $n = \sum_i n_i \rightarrow \infty$. We assume that $\mathbf{n} = np$ has a multinomial distribution with probabilities $\boldsymbol{\pi} = (\pi_1, \dots, \pi_N)^T$. In general terms, the model is

$$\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta}),$$

where $\boldsymbol{\pi}(\boldsymbol{\theta})$ denotes a function that relates $\boldsymbol{\pi}$ to a smaller number of parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q)^T$. We use $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ to denote generic parameter and probability values, and $\boldsymbol{\theta}_0 = (\theta_{10}, \dots, \theta_{q0})^T$ and $\boldsymbol{\pi}_0 = (\pi_{10}, \dots, \pi_{N0})^T = \boldsymbol{\pi}(\boldsymbol{\theta}_0)$ to denote true values for a particular application. When the model does not hold, no $\boldsymbol{\theta}_0$ exists for which $\boldsymbol{\pi}(\boldsymbol{\theta}_0) = \boldsymbol{\pi}_0$; that is, $\boldsymbol{\pi}_0$ falls outside the subset of $\boldsymbol{\pi}$ values that is the range of $\boldsymbol{\pi}(\boldsymbol{\theta})$ for the space of possible $\boldsymbol{\theta}$. We consider this case in Section 16.3.5.

We first derive the asymptotic distribution of the ML estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. We use that to derive the asymptotic distribution of the model-based ML estimator $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})$ of $\boldsymbol{\pi}$. The assumed regularity conditions are:

1. $\boldsymbol{\theta}_0$ is not on the boundary of the parameter space.
2. All $\pi_{i0} > 0$.
3. $\boldsymbol{\pi}(\boldsymbol{\theta})$ has continuous first-order partial derivatives in a neighborhood of $\boldsymbol{\theta}_0$.
4. The Jacobian matrix $(\partial \boldsymbol{\pi} / \partial \boldsymbol{\theta})$ has full rank q at $\boldsymbol{\theta}_0$.

These conditions ensure that $\boldsymbol{\pi}(\boldsymbol{\theta})$ is locally smooth and one-to-one at $\boldsymbol{\theta}_0$ and Taylor series expansions exist in neighborhoods around $\boldsymbol{\theta}_0$ and $\boldsymbol{\pi}_0$.

As in the Cramér (1946) and Rao (1957) proofs, the derivations regard the ML estimate as a point in the parameter space where the derivative of the log-likelihood function is zero. Birch (1964a) regarded it as a point at which the likelihood takes value arbitrarily near its supremum. Although his approach is more powerful, the proofs are more complex. In assuming that an ML estimator of $\boldsymbol{\theta}$ exists and is a solution of the likelihood equations, we require a *strong identifiability* condition: For every $\epsilon > 0$, there exists a $\delta > 0$ such that if $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \epsilon$, then $\|\boldsymbol{\pi}(\boldsymbol{\theta}) - \boldsymbol{\pi}_0\| > \delta$. This condition implies a weaker one that two $\boldsymbol{\theta}$ values cannot have the same $\boldsymbol{\pi}$ value. When strong identifiability and the other regularity conditions hold, the probability an ML estimator is a root of the likelihood equations converges to 1 as $n \rightarrow \infty$. That estimator then has the standard asymptotic properties of a solution of the likelihood equations. For proofs with slightly weaker regularity conditions, see Rao (1973, Sec. 5e) and Bishop et al. (1975, Secs. 14.7 and 14.8).

16.2.1 Asymptotic Distribution of Model Parameter Estimator

Suppose that observations are independent from $f(\mathbf{y}; \boldsymbol{\theta})$, some probability mass function. The ML estimator $\hat{\boldsymbol{\theta}}$ is efficient, in the sense that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\mathcal{J}}^{-1}),$$

where $\boldsymbol{\mathcal{J}}$ is the information matrix for a single observation. The (j, k) element of $\boldsymbol{\mathcal{J}}$ is

$$-E\left(\frac{\partial^2 \log f(\mathbf{y}, \boldsymbol{\theta})}{\partial \theta_j \partial \theta_k}\right) = E\left[\frac{\partial \log f(\mathbf{y}, \boldsymbol{\theta})}{\partial \theta_j} \cdot \frac{\partial \log f(\mathbf{y}, \boldsymbol{\theta})}{\partial \theta_k}\right].$$

When f is the probability of an observation having multinomial probabilities $(\pi_1(\boldsymbol{\theta}), \dots, \pi_N(\boldsymbol{\theta}))$, this element of $\boldsymbol{\mathcal{J}}$ equals

$$\sum_{i=1}^N \frac{\partial \log(\pi_i(\boldsymbol{\theta}))}{\partial \theta_j} \frac{\partial \log(\pi_i(\boldsymbol{\theta}))}{\partial \theta_k} \pi_i(\boldsymbol{\theta}) = \sum_{i=1}^N \frac{\partial \pi_i(\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \pi_i(\boldsymbol{\theta})}{\partial \theta_k} \frac{1}{\pi_i(\boldsymbol{\theta})}.$$

We'll express these elements more simply in matrix form. Let A denote the $N \times q$ matrix having elements

$$a_{ij} = \pi_{i0}^{-1/2} \frac{\partial \pi_i(\boldsymbol{\theta})}{\partial \theta_{j0}}.$$

The matrix expression for A is

$$(16.10) \quad A = \text{Diag}(\pi_0)^{-1/2} (\partial \boldsymbol{\pi} / \partial \boldsymbol{\theta}_0),$$

where $(\partial \boldsymbol{\pi} / \partial \boldsymbol{\theta}_0)$ denotes the Jacobian $(\partial \boldsymbol{\pi} / \partial \boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}_0$. So, the above element of $\boldsymbol{\mathcal{J}}$ equals the (j, k) element of $A^T A$. Since the Jacobian has full rank at $\boldsymbol{\theta}_0$, $A^T A$ is nonsingular. Thus,

$$(16.11) \quad \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, (A^T A)^{-1}).$$

The asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ depends on $(\partial \boldsymbol{\pi} / \partial \boldsymbol{\theta}_0)$ and hence on the function for modeling $\boldsymbol{\pi}$ in terms of $\boldsymbol{\theta}$.

16.2.2 Asymptotic Distribution of Cell Probability Estimators

The asymptotic distribution of the model-based estimator $\hat{\pi}$ follows from the Taylor series expansion

$$(16.12) \quad \hat{\pi} = \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\pi}(\boldsymbol{\theta}_0) + \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(n^{-1/2}).$$

The size of the remainder term follows from $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = O_p(n^{-1/2})$.

Now it $\boldsymbol{\pi}(\boldsymbol{\theta}_0) = \boldsymbol{\pi}_0$, and $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normal with asymptotic covariance $(\mathbf{A}^T \mathbf{A})^{-1}$. By the delta method,

$$(16.13) \quad \sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0) \xrightarrow{d} N\left[\mathbf{0}, \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}_0} (\mathbf{A}^T \mathbf{A})^{-1} \frac{\partial \boldsymbol{\pi}^T}{\partial \boldsymbol{\theta}_0}\right].$$

The marginal approximation for a particular probability that is very close to 0 may require quite large n to be good.

16.2.3 Model Smoothing Is Beneficial

When the model holds with $\boldsymbol{\theta}$ having $q < N - 1$ elements, $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})$ is more efficient than the sample proportion \mathbf{p} for estimating $\boldsymbol{\pi}$. More generally, for estimating a smooth function $g(\boldsymbol{\pi})$ of $\boldsymbol{\pi}$, $g(\hat{\boldsymbol{\pi}})$ has smaller asymptotic variance than $g(\mathbf{p})$. Altham (1984) proved this result. Her proof applies not only to categorical data but to any situation in which a model describes the dependence of a set of parameters on some smaller set. The proof uses standard properties of ML estimators and applies whenever regularity conditions hold that guarantee those properties.

Let $\Sigma = \text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T$ denote the covariance matrix of $\sqrt{n} \mathbf{p}$. By the delta method,

$$\text{asympt. var}[\sqrt{n} g(\mathbf{p})] = \left(\frac{\partial g}{\partial \boldsymbol{\pi}_0} \right)^T [\text{cov}(\sqrt{n} \mathbf{p})] \frac{\partial g}{\partial \boldsymbol{\pi}_0} = \left(\frac{\partial g}{\partial \boldsymbol{\pi}_0} \right)^T \boldsymbol{\Sigma} \frac{\partial g}{\partial \boldsymbol{\pi}_0}$$

and

$$\begin{aligned} \text{asympt. var}[\sqrt{n} g(\hat{\boldsymbol{\pi}})] &= \left(\frac{\partial g}{\partial \boldsymbol{\pi}_0} \right)^T [\text{asympt. cov}(\sqrt{n} \hat{\boldsymbol{\pi}})] \frac{\partial g}{\partial \boldsymbol{\pi}_0} \\ &= \left(\frac{\partial g}{\partial \boldsymbol{\pi}_0} \right)^T \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}_0} [\text{asympt. cov}(\sqrt{n} \hat{\boldsymbol{\theta}})] \left(\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}_0} \right)^T \frac{\partial g}{\partial \boldsymbol{\pi}_0}. \end{aligned}$$

From (16.10) and (16.11),

$$\text{asympt. cov}(\sqrt{n} \hat{\boldsymbol{\theta}}) = (\mathbf{A}^T \mathbf{A})^{-1} = [(\partial \boldsymbol{\pi} / \partial \boldsymbol{\theta}_0)^T \text{Diag}(\boldsymbol{\pi}_0)^{-1} (\partial \boldsymbol{\pi} / \partial \boldsymbol{\theta}_0)]^{-1}.$$

16.3 ASYMPTOTIC DISTRIBUTIONS OF RESIDUALS AND GOODNESS-OF-FIT STATISTICS

We next derive the distribution of the Pearson X^2 and likelihood-ratio G^2 goodness-of-fit statistics for a multinomial model $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$. We first derive the asymptotic joint distribution of the sample proportions \mathbf{p} and model-based estimator $\hat{\boldsymbol{\pi}}$. This distribution determines large-sample distributions of statistics that depend on both \mathbf{p} and $\hat{\boldsymbol{\pi}}$, such as residuals. Deriving the large-sample chi-squared distribution for X^2 , which is the sum of squared Pearson residuals, is then straightforward. We also show that X^2 and G^2 are asymptotically equivalent, when the model holds. The presentation borrows from Bishop et al. (1975, Chap. 14), Cox (1984), Cramér (1946, pp. 432–433), and Rao (1973, Sec. 6b).

16.3.1 Joint Asymptotic Normality of \boldsymbol{p} and $\hat{\boldsymbol{\pi}}$

We first express the joint dependence of \boldsymbol{p} and $\hat{\boldsymbol{\pi}}$ on \boldsymbol{p} , in order to show the joint asymptotic normality of \boldsymbol{p} and $\hat{\boldsymbol{\pi}}$. Let

$$\mathbf{D} = \text{Diag}(\boldsymbol{\pi}_0)^{1/2} \mathbf{A} (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \text{Diag}(\boldsymbol{\pi}_0)^{-1/2}.$$

From (16.11) and (16.12),

$$\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0 = \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}_0} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_p(n^{-1/2}) = \mathbf{D}(\boldsymbol{p} - \boldsymbol{\pi}_0) + o_p(n^{-1/2}).$$

Therefore,

$$\sqrt{n} \begin{pmatrix} \boldsymbol{p} - \boldsymbol{\pi}_0 \\ \hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0 \end{pmatrix} = \begin{pmatrix} \mathbf{I} \\ \mathbf{D} \end{pmatrix} \sqrt{n} (\boldsymbol{p} - \boldsymbol{\pi}_0) + o_p(1),$$

where \mathbf{I} is a $N \times N$ identity matrix. By the delta method,

$$(16.14) \quad \sqrt{n} \begin{pmatrix} \boldsymbol{p} - \boldsymbol{\pi}_0 \\ \hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0 \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}^*),$$

where

$$(16.15) \quad \boldsymbol{\Sigma}^* = \begin{pmatrix} \text{Diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0 \boldsymbol{\pi}_0^T & [\text{Diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0 \boldsymbol{\pi}_0^T] \mathbf{D}^T \\ \mathbf{D} [\text{Diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0 \boldsymbol{\pi}_0^T] & \mathbf{D} [\text{Diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0 \boldsymbol{\pi}_0^T] \mathbf{D}^T \end{pmatrix}.$$

The two matrix blocks on the main diagonal of $\boldsymbol{\Sigma}^*$ are $\text{cov}(\sqrt{n} \boldsymbol{p})$ and asymp. $\text{cov}(\sqrt{n} \hat{\boldsymbol{\pi}})$, derived previously. The new information here is that asymp. $\text{cov}(\sqrt{n} \boldsymbol{p}, \sqrt{n} \hat{\boldsymbol{\pi}}) = [\text{Diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0 \boldsymbol{\pi}_0^T] \mathbf{D}^T$.

16.3.2 Asymptotic Distribution of Pearson and Standardized Residuals

For cell counts $\{n_i\}$ the Pearson statistic is $X^2 = \sum_i e_i^2$, where

$$e_i = \frac{n_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}} = \frac{\sqrt{n}(p_i - \hat{\pi}_i)}{\sqrt{\hat{\pi}_i}}.$$

For Poisson models, this is the Pearson residual. The residuals e are functions of p and $\hat{\pi}$, which are jointly asymptotically normal from (16.15). To use the delta method, we calculate

$$\partial e_i / \partial p_i = \sqrt{n} \hat{\pi}_i^{-1/2}, \quad \partial e_i / \partial \hat{\pi}_i = -\sqrt{n} (p_i + \hat{\pi}_i) / 2\hat{\pi}_i^{-3/2},$$

$$\partial e_i / \partial p_j = \partial e_i / \partial \hat{\pi}_j = 0 \quad \text{for } i \neq j.$$

That is,

$$(16.16) \quad \begin{aligned} \frac{\partial \mathbf{e}}{\partial \mathbf{p}} &= \sqrt{n} \mathbf{Diag}(\hat{\pi})^{-1/2} \quad \text{and} \\ \frac{\partial \mathbf{e}}{\partial \hat{\pi}} &= -\left(\frac{1}{2}\right) \sqrt{n} [\mathbf{Diag}(p) + \mathbf{Diag}(\hat{\pi})] \mathbf{Diag}(\hat{\pi})^{-3/2}. \end{aligned}$$

Evaluated at $p = \pi_0$ and $\hat{\pi} = \hat{\pi}_0$, these matrices equal $\sqrt{n} \mathbf{Diag}(\pi_0)^{-1/2}$ and $-\sqrt{n} \mathbf{Diag}(\pi_0)^{-1/2}$. Using (16.16), (16.17), and $A^T \pi^{1/2} \mathbf{0} = \mathbf{0}$ [which follows from

$$\sum_i \partial \pi_i(\theta) / \partial \theta_j = \partial / \partial \theta_j [\sum_i \pi_i(\theta)] = \partial / \partial \theta_j(1) = 0],$$

$$(16.17) \quad \mathbf{e} \xrightarrow{d} N(\mathbf{0}, \mathbf{I} - \pi_0^{1/2} (\pi_0^{1/2})^T - A(A^T A)^{-1} A^T).$$

The limiting distribution has form $N(\mathbf{0}, \mathbf{I} - \mathbf{H}_{at})$, where \mathbf{H}_{at} is the *hat matrix* (Section 4.5.6). The standardized residual (Haberman 1973a) divides e by its estimated standard error. This statistic, which is asymptotically standard normal, equals

$$(16.18) \quad r_i = \frac{e_i}{[1 - \hat{\pi}_i - \sum_j \sum_k (1/\hat{\pi}_i)(\partial \pi_i / \partial \hat{\theta}_j)(\partial \pi_i / \partial \hat{\theta}_k) \hat{\nu}^{jk}]^{1/2}},$$

where $\hat{\nu}^{jk}$ denotes the element in row j and column k of $(\hat{\mathbf{A}}^T \hat{\mathbf{A}})^{-1}$. The denominator of r_i is $\sqrt{1 - \hat{h}_i}$, where the leverage \hat{h}_i for observation i estimates the i th diagonal element of the hat matrix.

16.3.3 Asymptotic Distribution of Pearson X^2 Statistic

The proof that the Pearson X^2 statistic has an asymptotic chi-squared distribution uses the following relationship between normal and chi-squared distributions (Rao 1973, p. 188):

Let \mathbf{X} be multivariate normal with mean $\boldsymbol{\nu}$ and covariance matrix \mathbf{B} . A necessary and sufficient condition for $(\mathbf{X} - \boldsymbol{\nu})^T \mathbf{C}(\mathbf{X} - \boldsymbol{\nu})$ to have a chi-squared distribution is $\mathbf{BCBCB} = \mathbf{BCB}$. The degrees of freedom equal the rank of \mathbf{CB} .

When \mathbf{B} is nonsingular, the condition simplifies to $\mathbf{CBC} = \mathbf{C}$.

The Pearson statistic relates to \mathbf{e} by $X^2 = \mathbf{e}^T \mathbf{e}$, so we apply this result by identifying \mathbf{X} with \mathbf{e} , $\boldsymbol{\nu} = \mathbf{0}$, $\mathbf{C} = \mathbf{I}$, and $\mathbf{B} = \mathbf{I} - \boldsymbol{\pi}_0^{1/2}(\boldsymbol{\pi}_0^{1/2})^T - \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$. Since $\mathbf{C} = \mathbf{I}$, the condition for $(\mathbf{X} - \boldsymbol{\nu})^T \mathbf{C}(\mathbf{X} - \boldsymbol{\nu}) = \mathbf{e}^T \mathbf{e} = X^2$ to have a chi-squared distribution simplifies to $\mathbf{BBB} = \mathbf{BB}$. A direct computation using $\mathbf{A}^T \boldsymbol{\pi}_0^{1/2} = \mathbf{0}$ shows that \mathbf{B} is idempotent, so the condition holds. Since \mathbf{e} is asymptotically multivariate normal, X^2 is asymptotically chi-squared.

For symmetric idempotent matrices, the rank equals the trace. The trace of \mathbf{I} is N ; the trace of $\boldsymbol{\pi}_0^{1/2}(\boldsymbol{\pi}_0^{1/2})^T$ equals the trace of $(\boldsymbol{\pi}_0^{1/2})^T \boldsymbol{\pi}_0^{1/2} = \sum_i \pi_{i0} = 1$, which is 1; the trace of $\mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$ equals the trace of $(\mathbf{A}^T \mathbf{A})^{-1}(\mathbf{A}^T \mathbf{A})$ = identity matrix of size $q \times q$, which is q . Thus, the rank of $\mathbf{B} = \mathbf{CB}$ is $N - q - 1$, and the asymptotic chi-squared distribution has $\text{df} = N - q - 1$.

The result, due to Fisher (1922), is remarkably simple. When the sample size is large, the distribution of X^2 does not depend on $\boldsymbol{\pi}_0$ or the model form. It depends only on the difference between the dimension of $\boldsymbol{\pi}$, which is $N - 1$, and the dimension of $\boldsymbol{\theta}$. Watson (1959) showed that the same result holds for the asymptotic conditional distribution, given a sufficient statistic for nuisance parameters. With $q = 0$ parameters, X^2 is Pearson's (1900) statistic (1.16) for testing that multinomial probabilities equal certain specified values. Then, $\text{df} = N - 1$, as Pearson claimed.

16.3.4 Asymptotic Distribution of Likelihood-Ratio Statistic

When the model holds, the likelihood-ratio statistic G^2 is asymptotically equivalent to X^2 as $n \rightarrow \infty$. To show this, we express

$$G^2 = 2 \sum_i n_i \log \frac{n_i}{\hat{\mu}_i} = 2n \sum_i p_i \log \left(1 + \frac{p_i - \hat{\pi}_i}{\hat{\pi}_i} \right)$$

and apply the expansion

$$\log(1+x) = x - x^2/2 + x^3/3 - \dots \quad \text{for } |x| < 1.$$

We identify x with $(p_i - \hat{\pi}_i)/\hat{\pi}_i$, which converges in probability to 0 when the model holds. For large n ,

$$\begin{aligned} G^2 &= 2n \sum_i [\hat{\pi}_i + (p_i - \hat{\pi}_i)] \left[\frac{p_i - \hat{\pi}_i}{\hat{\pi}_i} - \left(\frac{1}{2} \right) \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i^2} + \dots \right] \\ &= 2n \sum_i \left[(p_i - \hat{\pi}_i) - \left(\frac{1}{2} \right) \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i} + \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i} + O_p(p_i - \hat{\pi}_i)^3 \right] \\ &= n \sum_i \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i} + 2nO_p(n^{-3/2}) = X^2 + O_p(n^{-1/2}) = X^2 + o_p(1), \end{aligned}$$

since $\sum_i (p_i - \hat{\pi}_i) = 0$ and $(p_i - \hat{\pi}_i) = (p_i - \pi_i) - (\hat{\pi}_i - \pi_i)$, both of which are $O_p(n^{-1/2})$. Thus, when the model holds, $X^2 - G^2 \rightarrow_p 0$. As a consequence, G^2 , like X^2 , has an asymptotic chi-squared distribution with $\text{df} = N - q - 1$.

The parameter value that maximizes the likelihood is the one that minimizes G^2 . To show this, we let

$$G^2(\boldsymbol{\pi}; \mathbf{p}) = 2n \sum_i p_i \log(p_i/\pi_i).$$

The kernel of the multinomial log-likelihood is

$$\begin{aligned} L(\boldsymbol{\theta}) &= n \sum_i p_i \log \pi_i(\boldsymbol{\theta}) \\ &= -n \sum_i p_i \log \frac{p_i}{\pi_i(\boldsymbol{\theta})} + n \sum_i p_i \log p_i \\ &= -\left(\frac{1}{2} \right) G^2(\boldsymbol{\pi}(\boldsymbol{\theta}); \mathbf{p}) + n \sum_i p_i \log p_i. \end{aligned}$$

The second term in the last expression does not depend on $\boldsymbol{\theta}$, so maximizing $L(\boldsymbol{\theta})$ is equivalent to minimizing G^2 with respect to $\boldsymbol{\theta}$.

A fundamental result for G^2 concerns comparisons of nested models. For two models, with M_0 a special case of M_1 , let q_0 and q_1 denote the numbers of parameters and let $\{\hat{\pi}_{0i}\}$ and $\{\hat{\pi}_{1i}\}$ denote ML estimators of cell probabilities. Then

$$G^2(M_0) - G^2(M_1) = 2n \sum_i p_i \log(\hat{\pi}_{1i}/\hat{\pi}_{0i})$$

has the form of $-2(\log\text{-likelihood ratio})$ for testing that M_0 holds against the alternative that M_1 holds. Theory for likelihood-ratio tests suggests that when the simpler model holds, its asymptotic distribution is chi-squared with $\text{df} = q_1 - q_0$. For details, see Bishop et al. (1975, pp. 525-526), Haberman (1974a, p. 108), and Rao (1973, pp. 418-419). The statistic $X^2(M_0|M_1)$ in (4.40) with $v(\hat{\mu}_{0i}) = \hat{\mu}_{0i}$ is a quadratic approximation for the G^2 difference. Haberman (1977a) noted that these tests can perform well even for large, sparse tables, as long as $q_1 - q_0$ is small relative to the sample size and no expected frequency has larger order of magnitude than the others.

16.3.5 Asymptotic Noncentral Distributions

Results in this chapter assume that a certain parametric model holds. In practice, any unsaturated model almost surely does not hold perfectly. This is not problematic if we regard models merely as convenient approximations for reality. For instance, the ML estimator $\hat{\theta}$ converges to a value θ_0 that describes the best fit of the chosen model to reality. In this sense, inferences for θ give us information about a useful approximation for reality.

For goodness-of-fit statistics, a relevant distinction exists between limiting behavior when the model holds and when it does not hold. When the model holds, X^2 and G^2 have a limiting chi-squared distribution, and the difference between them disappears as n increases. When the model does not hold, X^2 and G^2 tend to grow unboundedly as n increases, and $|X^2 - G^2|$ need not go to zero. One method for obtaining proper limiting distributions considers a sequence of situations π_n for which the lack of fit diminishes as n increases. Specifically, the model is $\pi = f(\theta)$, but in reality

$$(16.19) \quad \pi_n = f(\theta) + \delta/\sqrt{n}.$$

The best fit of the model to the population has i th probability equal to $f_i(\theta)$, but the true value differs from that by δ_i/\sqrt{n} .

For this representation, Mitra (1958) showed that X^2 has a limiting noncentral chi-squared distribution, with $\text{df} = N - q - 1$ and noncentrality parameter

$$\lambda = n \sum_{i=1}^n \frac{[\pi_{ni} - f_i(\theta)]^2}{f_i(\theta)}.$$

This has the form of X^2 , with the sample values p_i and $\hat{\pi}_i$, replaced by population values π_{ni} and $f_i(\theta)$. Similarly, the noncentrality of the likelihood-ratio statistic has the form of G^2 , with the same substitution. Haberman (1974a, pp. 109–112) showed that under certain conditions G^2 and X^2 have the same limiting distribution; that is, their noncentrality values converge to a common value as $n \rightarrow \infty$.

Representation (16.20) means that, for large n , the noncentral chi-squared approximation is valid when the model is just barely incorrect. In practice, it is often reasonable to adopt (16.20) for fixed n to approximate the distribution of X_2 , even though (16.20) would not be plausible as we obtain more data. The alternative representation

$$(16.20) \quad \pi = f(\theta) + \delta$$

in which π differs from $f(\theta)$ by a *fixed* amount as $n \rightarrow \infty$ may seem more natural. In fact, this is more appropriate than (16.20) for proving the test to be consistent (i.e., for convergence to 1 of the probability of rejecting the hypothesis that the model holds). For (16.21), however, the noncentrality parameter λ grows unboundedly as $n \rightarrow \infty$, and a proper limiting distribution does not result for X^2 and G^2 .

When the model holds, $\delta = \mathbf{0}$ in either representation (16.20) or (16.21). That is, $f(\theta) = \pi(\theta)$, $\lambda = 0$, and the results in Sections 16.3.3 and 16.3.4 apply.

16.4 ASYMPTOTIC DISTRIBUTIONS FOR LOGIT/LOGLINEAR MODELS

For loglinear models, formulas in Section 9.6 for the asymptotic covariance matrices of θ and $\hat{\pi}$ are special cases of ones derived in Section 16.2. We present these for the multinomial form of the models, which relates directly to that section. Then we discuss the connection to Poisson loglinear models.

To constrain probabilities to sum to 1, we express loglinear models for multinomial sampling as

$$(16.21) \quad \pi = \exp(\mathbf{X}\boldsymbol{\theta}) / [\mathbf{1}^T \exp(\mathbf{X}\boldsymbol{\theta})],$$

where \mathbf{X} is a model matrix and $\mathbf{1}^T = (1, \dots, 1)$. Letting \mathbf{x}_i denote row i of \mathbf{X} ,

$$\pi_i = \pi_i(\boldsymbol{\theta}) = \frac{\exp(\mathbf{x}_i\boldsymbol{\theta})}{\sum_k \exp(\mathbf{x}_k\boldsymbol{\theta})}.$$

16.4.1 Asymptotic Covariance Matrices

A model affects covariance matrices through the Jacobian. Since

$$\begin{aligned}\frac{\partial \pi_i}{\partial \theta_j} &= \frac{[\sum_k \exp(\mathbf{x}_k \boldsymbol{\theta})][\exp(\mathbf{x}_i \boldsymbol{\theta})]x_{ij} - [\exp(\mathbf{x}_i \boldsymbol{\theta})][\sum_k x_{kj} \exp(\mathbf{x}_k \boldsymbol{\theta})]}{\sum_k \exp(\mathbf{x}_k \boldsymbol{\theta})^2} \\ &= \pi_i x_{ij} - \pi_i \sum_k x_{kj} \pi_k,\end{aligned}$$

the matrix of these elements has the form

$$\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}} = [\text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi} \boldsymbol{\pi}^T] \mathbf{X}.$$

Using this with (16.10) and (16.11), the information matrix at $\boldsymbol{\theta}_0$ is

$$\begin{aligned}\mathbf{A}^T \mathbf{A} &= (\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}_0})^T \text{Diag}(\boldsymbol{\pi}_0)^{-1} (\frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\theta}_0}) \\ &= \mathbf{X}^T [\text{Diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0 \boldsymbol{\pi}_0^T]^T \text{Diag}(\boldsymbol{\pi}_0)^{-1} [\text{Diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0 \boldsymbol{\pi}_0^T] \mathbf{X} \\ &= \mathbf{X}^T [\text{Diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0 \boldsymbol{\pi}_0^T] \mathbf{X}.\end{aligned}$$

Thus, for multinomial loglinear models, $\hat{\boldsymbol{\theta}}$ is asymptotically normally distributed with estimated covariance matrix

$$(16.22) \quad \widehat{\text{cov}}(\hat{\boldsymbol{\theta}}) = \{ \mathbf{X}^T [\text{Diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}} \hat{\boldsymbol{\pi}}^T] \mathbf{X} \}^{-1} / n.$$

Similarly, from (16.13) the estimated asymptotic covariance matrix of $\hat{\boldsymbol{\pi}}$ is

$$\widehat{\text{cov}}(\hat{\boldsymbol{\pi}}) = \frac{1}{n} [\text{Diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}} \hat{\boldsymbol{\pi}}^T] \mathbf{X} \{ \mathbf{X}^T [\text{Diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}} \hat{\boldsymbol{\pi}}^T] \mathbf{X} \}^{-1} \mathbf{X}^T [\text{Diag}(\hat{\boldsymbol{\pi}}) - \hat{\boldsymbol{\pi}} \hat{\boldsymbol{\pi}}^T].$$

From (16.18), the Pearson residuals \mathbf{e} are asymptotically normal with

$$\begin{aligned}\text{asympt. cov}(\mathbf{e}) &= \mathbf{I} - \boldsymbol{\pi}_0^{1/2} \left(\boldsymbol{\pi}_0^{1/2} \right)^T - \mathbf{A} (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \\ &= \mathbf{I} - \boldsymbol{\pi}_0^{1/2} \left(\boldsymbol{\pi}_0^{1/2} \right)^T - \text{Diag}(\boldsymbol{\pi}_0)^{-1/2} [\text{Diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0 \boldsymbol{\pi}_0^T] \mathbf{X} \\ &\quad \times \{ \mathbf{X}^T [\text{Diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0 \boldsymbol{\pi}_0^T] \mathbf{X} \}^{-1} \\ &\quad \times \mathbf{X}^T [\text{Diag}(\boldsymbol{\pi}_0) - \boldsymbol{\pi}_0 \boldsymbol{\pi}_0^T] \text{Diag}(\boldsymbol{\pi}_0)^{-1/2}.\end{aligned}$$

16.4.2 Connection with Poisson Loglinear Models

This book expresses loglinear models in terms of Poisson expected cell frequencies $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)^T$, using formulas of the form

$$(16.23) \log \boldsymbol{\mu} = \mathbf{X}_a \boldsymbol{\theta}_a.$$

The model matrix \mathbf{X}_a and parameter vector $\boldsymbol{\theta}_a$ in this formula are slightly different from \mathbf{X} and $\boldsymbol{\theta}$ in multinomial model (16.22). The Poisson expression (16.24) does not have constraints on $\boldsymbol{\mu}$. For multinomial model (16.22), $\sum_i \mu_i = n$ is fixed, and $\boldsymbol{\pi} = \boldsymbol{\mu}/n$ satisfies

$$\begin{aligned} \log \boldsymbol{\mu} &= \log n \boldsymbol{\pi} = \mathbf{X}\boldsymbol{\theta} + [\log n - \log(\mathbf{1}^T \exp(\mathbf{X}\boldsymbol{\theta}))] \mathbf{1} \\ &= \mathbf{X}\boldsymbol{\theta} + \mathbf{1}\lambda, \end{aligned}$$

where $\lambda = \log n - \log(\mathbf{1}^T \exp(\mathbf{X}\boldsymbol{\theta}))$. In other words, multinomial model (16.22) implies Poisson model (16.24) with

$$\mathbf{X}_a = [\mathbf{1}; \mathbf{X}] \quad \text{and} \quad \boldsymbol{\theta}_a = (\lambda, \boldsymbol{\theta}^T)^T.$$

The columns of \mathbf{X} in the multinomial representation must be linearly independent of $\mathbf{1}$; that is, the parameter λ , which relates to the total sample size, does not appear in $\boldsymbol{\theta}$. The dimension of $\boldsymbol{\theta}$ is 1 less than the number of parameters reported in this text for Poisson loglinear models. For instance, for the saturated model, $\boldsymbol{\theta}$ has $N - 1$ elements for the multinomial representation, reflecting the sole constraint on $\boldsymbol{\pi}$ of $\sum_i \pi_i = 1$.

16.5 SMALL-SAMPLE SIGNIFICANCE TESTS FOR CONTINGENCY TABLES

With modern computational power, it is not necessary to rely on large-sample approximations when n is small or when there is a large number of parameters. For many cases, tests and confidence intervals can directly use small-sample distributions rather than normal and chi-squared approximations. We studied small-sample methods in Sections 3.5 and 7.3, such as Fisher's exact test for testing independence. We next address this more generally for inference in contingency tables.¹

16.5.1 Exact Conditional Distribution for $I \times J$ Tables Under Independence

We first derive the distribution used in exact conditional tests of independence for $I \times J$ tables. We assume independent multinomial sampling within rows, as often applies in comparing I treatment groups. Then row totals $\{n_{i+}\}$ are fixed, and we estimate the I conditional distributions $\{\pi_{j|i}, j = 1, \dots, J\}$. Under H_0 : independence (i.e., homogeneity), $\pi_{j|1} = \pi_{j|2} = \dots = \pi_{j|I} = \pi_{+j}$, for $j = 1, \dots, J$. The product of the I multinomial probability functions then simplifies to

$$(16.24) \quad \prod_i \left(\frac{n_{i+}!}{\prod_j n_{ij}!} \prod_j \pi_{j|i}^{n_{ij}} \right) = \frac{(\prod_i n_{i+}!) (\prod_j \pi_{+j}^{n_{+j}})}{\prod_i \prod_j n_{ij}!}.$$

This distribution for $\{n_{ij}\}$ depends on $\{\pi_{+j}\}$. These are nuisance parameters, since they do not describe the association. Fisher proposed eliminating nuisance parameters by conditioning on their sufficient statistics. From the definition of sufficiency, the resulting conditional distribution does not depend on those parameters.

The contribution of $\{\pi_{+j}\}$ to the product multinomial distribution (16.25) depends on the data only through $\{n_{+j}\}$, which are their sufficient statistics. The $\{n_{+j}\}$ have the multinomial $(n, \{\pi_{+j}\})$ distribution, namely,

$$(16.25) \quad \frac{n!}{\prod_j n_{+j}!} \prod_j \pi_{+j}^{n_{+j}}.$$

The joint probability function of $\{n_{ij}\}$ and $\{n_{+j}\}$ is identical to the probability function of $\{n_{ij}\}$, since $\{n_{ij}\}$ determines $\{n_{+j}\}$. Thus, the probability function of $\{n_{ij}\}$, conditional on $\{n_{+j}\}$, equals the probability function (16.25) of $\{n_{ij}\}$ divided by the probability function (16.26) evaluated at $\{n_{+j}\}$. This gives cell probabilities

$$(16.26) \quad p(\{n_{ij}\} | \{n_{i+}\}, \{n_{+j}\}) = \frac{(\prod_i n_{i+}!) (\prod_j n_{+j}!)}{n! \prod_i \prod_j n_{ij}!}.$$

This *multivariate hypergeometric* distribution applies to the set of $\{n_{ij}\}$ having the same $\{n_{i+}\}$ and $\{n_{+j}\}$ as the observed table. For 2×2 tables, it is the hypergeometric distribution (3.17). When a table has a single multinomial sample, the unknown parameters are $\{\pi_{ij}\}$. For testing independence ($\pi_{ij} = \pi_{i+} \pi_{+j}$ all i and j), distribution (16.27) results from conditioning on the row and column totals. These are sufficient statistics for $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$, which determine the null distribution. For either sampling model, both sets of margins are fixed after the conditioning. The end result (16.27) does not depend on unknown parameters and thus permits exact probability calculations.

16.5.2 Exact Tests of Independence for $I \times J$ Tables

Exact tests of independence for $I \times J$ tables utilize the multivariate hypergeometric distribution. Freeman and Halton (1951) defined the P -value as the probability of the set of tables with the given margins that are no more likely to occur than the table observed. Other exact tests order the tables using a statistic describing distance from H_0 . Yates (1934) used X^2 . The P -value is then the null value of $P(X^2 \geq X^2_o)$ for observed value X^2_o , that is, the sum of the multivariate hypergeometric probabilities (16.27) for all tables with the given margins that have X^2 at least as large as observed. When classifications have ordered categories, an ordinal statistic is more relevant. For the one-sided alternative hypothesis of a positive association, we could use $P(T \geq t_o)$, where T is the correlation or gamma and t_o is its observed value.

Algorithms and software for exact tests for $I \times J$ tables are widely available (e.g., Mehta and Patel 1983). We recommend these tests when asymptotic approximations may be invalid. Computing time increases exponentially as n , I , or J increase. However, we can use Monte Carlo to sample randomly from the set of tables with the given margins (Agresti et al. 1979). The estimated P -value is then the sample proportion of tables having test statistic value at least as large as the value observed.

As I and/or J increase, the number of possible values for any test statistic T tends to increase. Thus, the conservativeness issue for conditional tests discussed in Section 3.5.5 becomes less problematic.

16.5.3 Example: Sexual Orientation and Party ID

We illustrate an exact test with [Table 16.1](#), which cross-classifies sexual orientation with political party ID for subjects of age 18–35 in the 2010 GSS. The first two rows contain many small counts, and large-sample tests of independence may be inappropriate. With small counts, the chi-squared approximation tends to be better with X^2 than G^2 . The value $X^2 = 19.18$ ($df = 12$) has P -value of 0.084.

Table 16.1 Sexual Orientation by Political Party ID

Sexual Orientation	Political Party ID						
	Strong Dem.	Dem.	Indep. near Dem.	Indep.	Indep. near Repub.	Repub.	Strong Repub.
Homosexual	1	3	3	0	0	1	0
Bisexual	4	2	2	7	0	0	0
Heterosexual	59	109	78	105	55	75	29

Source: 2010 General Social Survey.

Conditional on both sets of margins, using X^2 as the test criterion, the null probability of the observed table and the more extreme tables [based on formula [\(16.27\)](#)] equals 0.080. So, in this case the large-sample test performed fine.

Alternatively, treating rows and columns as ordinal, we could use an ordinal statistic to order the tables, potentially giving greater power and permitting one-sided tests. Using the correlation with row and column numbers as the scores, the sample correlation is 0.095. The exact P -value is 0.030 for the two-sided alternative and 0.014 for the negative association alternative, suggesting that heterosexuals are more likely to be Republican. The evidence is stronger than using X^2 , which ignores the ordering of categories.

16.6 SMALL-SAMPLE CONFIDENCE INTERVALS FOR CATEGORICAL DATA

We next consider small-sample interval estimation. For a given test about a particular parameter θ , a $100(1 - \alpha)\%$ test-based confidence interval (CI) for θ consists of all θ_0 for which P -values exceed α in the test of $H_0: \theta = \theta_0$.

16.6.1 Small-Sample CIs for a Binomial Parameter

We first consider a binomial parameter π . In Section 1.4.4 we tested $H_0: \pi = \pi_0$ directly using the binomial distribution. The best known small-sample interval, proposed by Clopper and Pearson (1934), uses the tail method for forming confidence intervals. It consists of all π_0 values for which each one-sided exact binomial P -value exceeds $\alpha/2$. With binomial outcome $Y = y$ in n trials, the lower and upper endpoints are the solutions in π_0 to the equations

$$\sum_{k=y}^n \binom{n}{k} \pi_0^k (1 - \pi_0)^{n-k} = \alpha/2 \quad \text{and} \quad \sum_{k=0}^y \binom{n}{k} \pi_0^k (1 - \pi_0)^{n-k} = \alpha/2,$$

except that the lower bound is 0 when $y = 0$ and the upper bound is 1 when $y = n$. When $y = 1, 2, \dots, n - 1$, from connections between binomial sums and the incomplete beta function and related cdf's of beta and F distributions, the confidence interval is

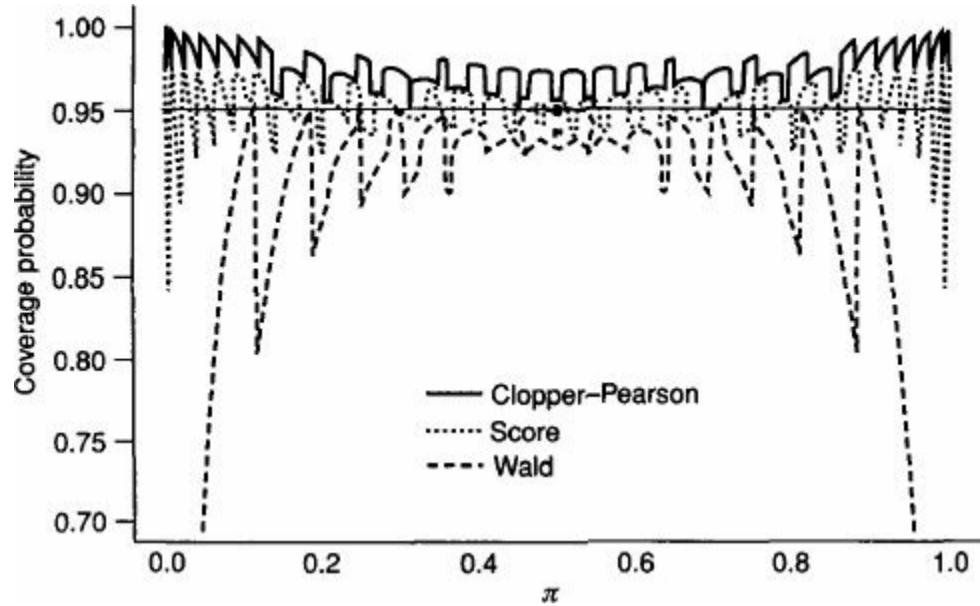
$$\left[1 + \frac{n - y + 1}{y F_{2y, 2(n-y+1)}(1 - \alpha/2)} \right]^{-1} < \pi < \left[1 + \frac{n - y}{(y + 1) F_{2(y+1), 2(n-y)}(\alpha/2)} \right]^{-1},$$

where $F_{a,b}(c)$ denotes the $1 - c$ quantile from the F distribution with $df_1 = a$ and $df_2 = b$. This interval corresponds to inverting a binomial two-sided test for which the P -value is double the minimum of the one-sided P -values.

In principle this approach seems ideal. However, there is a complication: Because of discreteness, the actual probability that the confidence interval contains the value of π is $\geq (1 - \alpha)$ rather than exactly $(1 - \alpha)$ (Neyman 1935, Casella and Berger 2001, p. 434). Similarly, for a test of $H_0: \pi = \pi_0$ at a fixed desired size α such as 0.05, it is not usually possible to achieve that size. With a finite number of possible samples, there is a finite number of possible P -values, of which 0.05 may not be one. In testing H_0 with fixed π_0 , we can pick a particular α that can occur as a P -value. For interval estimation, however, this is not an option. This is because constructing the interval corresponds to inverting an entire range of π_0 values in $H_0: \pi = \pi_0$, and each distinct π_0 value can have its own set of possible P -values; that is, there is not a single null parameter value π_0 as in one test.

The actual coverage probability can be much larger than the nominal confidence level. When $n = 25$, [Figure 16.1](#) plots the coverage probabilities as a function of the true parameter value π , for the Clopper-Pearson method, the large-sample score method, and the Wald method. At a fixed π value with a given method, the coverage probability is the sum of the binomial probabilities of all those samples for which the resulting interval contains that π . With $n = 25$, there are 26 possible samples and 26 corresponding confidence intervals, so the coverage probability is a sum of somewhere between 0 and 26 binomial probabilities. As π moves from 0 to 1, this coverage probability jumps up or down whenever π moves into or out of one of these 26 intervals. [Figure 16.1](#) shows that coverage probabilities are too low for the Wald method, whereas the Clopper-Pearson method errs in the opposite direction. The score method behaves well, its coverage probabilities tending to be near the nominal level, except for some π values close to 0 or 1. This is a good method even with relatively small n , unless π is near 0 or 1 (see Exercise 16.32).

[Figure 16.1](#) Plot of coverage probabilities for nominal 95% confidence intervals for binomial parameter π when $n = 25$.



In discrete problems using small-sample distributions, shorter confidence intervals result from inverting a single two-sided test rather than two one-sided tests as the Clopper-Pearson method does. For the binomial parameter, see Sterne (1954), Blyth and Still (1983), and Blaker (2000) for methods, summarized by Agresti and Min (2001) and Fay (2010a,b).

With the Sterne approach, for observed outcome y_o , the test of $H_0: \pi = \pi_0$ has P -value that sums up $P_{\pi_0}(y)$ for all outcomes y with $P_{\pi_0}(y) \leq P_{\pi_0}(y_o)$. This leads to optimality in terms of minimizing total length. In letters to the editor of *Applied Statistics*, Mantel and Halperin criticized this method in 1981 (pp. 73–74), but Barnard criticized their criticism in 1982 (pp. 304–305). With Blaker's approach the P -value is the minimum one-tail probability plus an attainable probability in the other tail that is as close as possible to, but not greater than, that one-tailed probability. This can be expressed as $P[Q_{\pi_0}(Y) \leq Q_{\pi_0}(y_o)]$ with $Q_{\pi_0}(y) = \min[P_{\pi_0}(Y \geq y), P_{\pi_0}(Y \leq y)]$. Its P -value cannot be greater than the Clopper-Pearson P -value. Thus, the corresponding interval is contained in the Clopper-Pearson interval and is preferable to it. The Blaker and Sterne intervals both have the nestedness property that an interval with larger confidence level necessarily contains one with a smaller level. However, they have inconsistencies, such as for certain data configurations having P -value that increases when an observation is added to the data set, regardless of its value (Fay 2010a, Vos and Hudson 2008).

16.6.2 CIs Based on Tests Using the Mid P -Value

In Section 1.4.4 we adjusted for discreteness in small-sample distributions by basing inference on the *mid P-value*. For a statistic T with observed result t_o for which larger results more strongly contradict H_0 , this is $\frac{1}{2} P(T = t_o) + P(T > t_o)$, less than the ordinary P -value of $P(T \geq t_o)$. Less conservative confidence intervals invert tests using the exact distribution with a mid P -value.

As with the ordinary P -value, there are various ways we could construct the interval. We'll illustrate for the binomial parameter. First, we could mimic the Clopper-Pearson construction (16.6.1) but replace each tail sum by the corresponding one-sided mid- P tail sum. This corresponds to inverting a test such that the 95% confidence interval is the set of π values for which double the minimum of the one-sided mid P -values exceeds 0.05, and it performs well while tending to be slightly conservative (Agresti and Gottard 2007). Brown et al. (2001) showed that this interval is similar to the Bayesian posterior interval generated with the Jeffreys prior distribution (beta with parameters 0.5 and 0.5). That interval has actual coverage probability close to the nominal level.

Another possible approach mimics the Blaker approach. It inverts the test for which the P -value is the minimum one-sided mid P -value plus the mid- P probability in the other tail that is as close as possible to that but no greater than it. An approach that mimics the Sterne interval inverts the test for which the P -value is the sum of $P_{\pi 0}(y)$ for all outcomes y with $P_{\pi 0}(y) < P_{\pi 0}(y_o)$ added to half the probability of y such that $P_{\pi 0}(y) = P_{\pi 0}(y_o)$. Yet another approach would use the ordinary mid P -value with a statistic T , such as $T = z^2$ for the score statistic (1.11).

16.6.3 Example: Proportion of Vegetarians Revisited

In Section 1.4.3 we estimated the proportion π of vegetarians in a population for which a sample of size $n = 25$ had $y = 0$ vegetarians. The 95% confidence intervals from inverting large-sample tests were $(0,0)$ for the Wald test, $(0,0.074)$ for the likelihood-ratio test, and $(0,0.133)$ for the score test.

For comparison, the Clopper–Pearson 95% interval for π is $(0.0, 0.137)$. This means that if we tested $H_0: \pi = 0.137$ against $H_a: \pi < 0.137$ and observed $y = 0$ in $n = 25$ trials, the binomial P -value = $P(Y \leq 0) = (1 - 0.137)^{25} = 0.025$. The 95% interval using the Blaker method is $(0.0, 0.128)$. The 95% interval using the mid- P adaptation of the Clopper-Pearson method is $(0.0, 0.113)$, which means that $\frac{1}{2} P(Y \leq 0) = \frac{1}{2}(1 - 0.113)^{25} = 0.025$.

16.6.4 Small-Sample CIs for Odds Ratios

To construct a small-sample confidence interval for the odds ratio, we can use a nonnull exact conditional distribution. For multinomial sampling, the distribution of $\{n_{ij}\}$ depends on n and cell probabilities $\{\pi_{ij}\}$. For 2×2 tables, the odds ratio is

$$\theta = \frac{\pi_{11} \pi_{22}}{\pi_{12} \pi_{21}} = \frac{\pi_{11}(1 - \pi_{1+} - \pi_{+1} + \pi_{11})}{(\pi_{1+} - \pi_{11})(\pi_{+1} - \pi_{11})}.$$

Hence, π_{11} is a function of θ and $\{\pi_{1+}, \pi_{+1}\}$. The same argument applies to any π_{ij} , so the multinomial distribution of $\{n_{ij}\}$ can use parameters $\{\theta, \pi_{1+}, \pi_{+1}\}$. Conditional on (n_{1+}, n_{+1}) , the distribution of $\{n_{ij}\}$ depends only on θ . Since n_{11} determines all other cell counts, given the marginal totals, the conditional distribution of $\{n_{ij}\}$ is specified by some function $P(n_{11} = t) = f(t; n_{1+}, n_{+1}, n, \theta)$. This distribution is the noncentral hypergeometric introduced in [\(7.9\)](#),

$$f(t|n_{1+}, n_{+1}, n; \theta) = \frac{\binom{n_{1+}}{t} \binom{n - n_{1+}}{n_{+1} - t} \theta^t}{\sum_{u=m_-}^{m_+} \binom{n_{1+}}{u} \binom{n - n_{1+}}{n_{+1} - u} \theta^u}. \quad (16.27)$$

The *conditional ML estimate* of θ is the value of θ that maximizes probability [\(16.28\)](#). Differentiating the log-likelihood with respect to θ shows that this estimate satisfies the equation $n_{11} = E(n_{11})$ in θ , where the expectation refers to distribution [\(16.28\)](#). This equation has a unique solution $\hat{\theta}$ and is solved using iterative methods (Cornfield 1956). This estimator differs from the *unconditional ML estimator* $\hat{\theta} = n_{11}n_{22}/n_{12}n_{21}$, which uses the ML estimates of $\{\pi_{ij}\}$ for the multinomial distribution of $\{n_{ij}\}$.

A confidence interval for θ results from inverting the test of $H_0: \theta = \theta_0$, having observed $n_{11} = t_o$. For $H_a: \theta > \theta_0$ and for $H_a: \theta < \theta_0$, the P -values are

$$P = \sum_{t \geq t_o} f(t; n_{1+}, n_{+1}, n, \theta_0) \quad \text{and} \quad P = \sum_{t \leq t_o} f(t; n_{1+}, n_{+1}, n, \theta_0).$$

When $\theta_0 = 1$, these are one-sided Fisher's exact tests. Mimicking the Clopper-Pearson approach, Cornfield (1956) set the lower endpoint as θ_0 for which $P = \alpha/2$ in testing against $H_a: \theta > \theta_0$ and the upper endpoint as θ_0 for which $P = \alpha/2$ for $H_a: \theta < \theta_0$. The interval is the set of θ_0 for which both one-sided P -values $\geq \alpha/2$.

As in Fisher's exact test, the conditional approach to interval estimation is necessarily conservative because of discreteness. The actual confidence coefficient, defined as the infimum of the coverage probabilities for all possible θ , has the nominal confidence level as a lower bound. Less conservative behavior and shorter intervals result from inverting a single two-sided test rather than inverting two one-sided tests. For the test criterion we could use the chi-squared score statistic for testing $H_0: \theta = \theta_0$, but utilize the exact conditional distribution to obtain the P -value (Agresti and Min 2001). Or, we could invert the test that has P -value equal to the minimum one-tail probability plus an attainable probability in the other tail that is as close as possible to, but not greater than, that one-tailed probability (Blaker 2000, Fay 2010). Or, we could invert the test that has P -value that sums the probabilities that are no greater than the probability of the observed table (Baptista and Pike 1977).

Using instead mid P -values to invert hypergeometric tests of $\theta = \theta_0$ yields narrower intervals with coverage probability usually nearer the nominal level, but not having that level as a lower bound. For interval estimation of the odds ratio, this method tends to be a bit conservative, but for small samples can yield much shorter intervals than Cornfield's exact interval.

An alternative approach with independent binomial samples inverts nonnull unconditional small-sample tests (Agresti and Min 2002, Lin and Yang 2006, Troendle and Frank 2001), an approach discussed in Section 16.6.8. Because of the reduced discreteness, such intervals are also usually shorter.

16.6.5 Example: Fisher's Tea Taster Revisited

We illustrate with [Table 3.9](#) from Fisher's tea-tasting experiment, for which we illustrated Fisher's exact test in Section 3.5.1. The conditional ML estimate of θ is 6.41. Software provides the Cornfield tail-method interval (0.21,626.24) with confidence coefficient guaranteed ≥ 0.95 . Not surprisingly, it is very wide because of the small sample. Inverting the family of two-sided exact conditional score tests gives a more precise interval, (0.31, 306.24). The unconditional approach is not appropriate here because of the sampling design.

Large-sample methods do not have the guarantee of bounds on error probabilities. They can be conservative or liberal, and thus their results can appear quite different from exact methods. For example, for the tea-tasting data, the 95% large-sample Wald confidence interval [\(3.2\)](#) for the odds ratio is (0.37,220.93), the large-sample score interval (proposed by Cornfield 1956 and by Miettinen and Nurminen 1985) without a continuity correction is (0.48,168.87), and the large-sample profile likelihood interval is (0.48,418.98). Normally, we would prefer an exact method over an approximate one. When the conditional distribution is highly discrete, however, the choice is not so obvious. Exact methods then can be quite conservative, especially with small samples.

For highly discrete data, it seems sensible to use adjustments of exact methods based on the mid P -value. For the tea-tasting data, for instance, the 95% confidence interval based on inverting two one-sided hypergeometric tests using the mid P -value is (0.31, 308.6), similar to the interval obtained by inverting the small-sample score test.

16.6.6 Small-Sample CIs for Logistic Regression Parameters

In Section 7.3 we used conditional ML to eliminate nuisance parameters in logistic regression models, by conditioning on their sufficient statistics. We used this approach also in Section 11.2 for matched-pairs data and in more general contexts for clustered data in Section 13.1. With the conditional likelihood we can use either large-sample methods or small-sample exact distributions. With the latter, we can conduct exact inference for logistic regression parameters, as explained in Section 7.3.2 and illustrated for Fisher’s exact test for 2×2 tables in Section 7.3.3, a test for the effect parameter in a linear logit model in Section 7.3.4, and a test of conditional independence in multiple 2×2 tables in Section 7.3.5. For that multiple 2×2 table case, we now illustrate small-sample confidence intervals.

For logistic model (7.11) of homogeneous association in $2 \times 2 \times K$ tables, the ordinary ML estimator of the odds ratio $\theta = \exp(\beta)$ behaves poorly for sparse-data asymptotics. The conditional ML estimator maximizes the conditional likelihood function after reducing the parameter space by conditioning on sufficient statistics for the other parameters (Andersen 1970, Birch 1964b). For cell counts $\{n_{ijk}\}$, given $\{n_{i+k}, n_{+jk}\}$ for all k , the conditional probability mass function that ($n_{111} = t_1, \dots, n_{11K} = t_K$) is the product of the hypergeometric functions (16.28) from the separate strata, or

$$(16.28) \quad \prod_k P(n_{11k} = t_k | n_{1+k}, n_{+1k}, n_{++k}; \theta) = \prod_k \frac{\binom{n_{1+k}}{t_k} \binom{n_{++k} - n_{1+k}}{n_{+1k} - t_k} \theta^{t_k}}{\sum_u \binom{n_{1+k}}{u} \binom{n_{++k} - n_{1+k}}{n_{+1k} - u} \theta^u}.$$

The conditional ML estimator $\hat{\theta}$ maximizes (16.29). Like the Mantel–Haenszel estimator $\hat{\theta}_{\text{MH}}$, it has good properties for both standard and sparse-data asymptotic cases (Andersen 1970, Breslow 1981), since the number of parameters does not change as K does. It can be slightly more efficient than $\hat{\theta}_{\text{MH}}$ except when $\theta = 1.0$, where they are equally efficient, or for matched pairs, where they are identical (Breslow 1981).

The conditional distribution (16.29) propagates one for $\sum_k n_{11k}$, which is used to test $H_0: \theta = \theta_0$ for an arbitrary value. Then, a 95% confidence interval for θ consists of all θ_0 for which the P -value exceeds 0.05. Such an interval is guaranteed to have at least the nominal coverage probability (Gart 1970; Kim and Agresti 1995, Mehta et al. 1985). This extends the interval for a single 2×2 table presented in Section 16.6.4.

Consider the promotion discrimination case in Table 7.5. There, $\sum_k n_{11k} = 0$, so the lower bound of any confidence interval for θ should be 0. For the generalization to several strata of Cornfield’s tail-method interval, StatXact software reports a 95% confidence interval of (0, 1.01). Using mid- P -values or P -values based on a finer partitioning of the sample space in tests and related confidence intervals reduces conservativeness (Note 3.10). Inverting exact tests of $H_0: \theta = \theta_0$ with the mid P -value yields the interval (0, 0.78). However, this approach cannot guarantee that the actual coverage probability is bounded below by 0.95.

Zelen (1971) presented a small-sample test of homogeneity of the odds ratios. Agresti (1992) discussed this and other small-sample methods for contingency tables. The methods of this section extend to estimating conditional odds ratios in models with several predictors, as the following example illustrates.

16.6.7 Example: Diarrhea and an Antibiotic

[Table 16.2](#) refers to 2493 patients having stays in a hospital. The response is whether they suffered an acute form of diarrhea during their stay. The three predictors are age (1 for over 50 years old, 0 for under 50), length of stay in hospital (1 for more than 1 week, 0 for less than 1 week), and exposure to an antibiotic called Cephalexin (1 for yes, 0 for no). We discuss estimation of the effect of Cephalexin, adjusting for age and length of stay, using a logistic model containing only main-effect terms.

Table 16.2 Data for Effect of Cephalexin Antibiotic on Diarrhea

Cephalexin ^a	Age ^a	Length of Stay ^a	Cases of Diarrhea	Sample Size
0	0	0	0	385
0	0	1	5	233
0	1	0	3	789
0	1	1	47	1081
1	1	1	5	5

^aSee the text for an explanation of 0 and 1.

Source: Based on study by E. Jaffe and V. Chang, Cornell Medical Center, reported in the Manual for *LogXact 7*. Cambridge, MA: CYTEL Software, 2005, p. 470.

The sample size is large, yet relatively few cases of acute diarrhea occurred. Moreover, all subjects having exposure to Cephalexin were also diarrhea cases, which causes an ML estimate of ∞ for the Cephalexin log odds ratio effect. To study that effect, we use an exact distribution, conditioning on sufficient statistics for the other predictors. Constructing a confidence interval by inverting the conditional test for the parameter, we obtain a 95% confidence interval of $(19, \infty)$ for the odds ratio. Assuming that the main-effects model is valid, Cephalexin appears to have a strong effect.

Results must be qualified somewhat because no Cephalexin cases occurred at the first three combinations of levels of age and length of stay. In fact, the first three rows of [Table 16.2](#) make no contribution to the analysis (Exercise 16.4). The data actually provide evidence about the effect of Cephalexin only for older subjects having a long stay.

16.6.8 Unconditional Small-Sample CIs for Difference of Proportions

The *conditional* approach to eliminating nuisance parameters works for parameters that have sufficient statistics. However, reduced sufficient statistics occur only for models that use the canonical parameters for the exponential family representation (e.g., logit for binomial, log mean for Poisson). For binary data, such models must be in terms of the log odds. For 2×2 tables, the conditional approach can yield confidence intervals for the log odds ratio but not for differences or ratios of proportions. An *unconditional* approach is more complex but does not require sufficient statistics.

Consider interval estimation of the difference of proportions for independent binomial samples. We used the unconditional approach in Section 3.5.6 for small-sample testing of $\pi_1 - \pi_2 = 0$. An unconditional confidence interval inverts the corresponding test of $H_0: \pi_1 - \pi_2 = \delta_0$, for any fixed $-1 < \delta_0 < 1$. The probability function for the table is the product of $\text{bin}(n_1, \pi_1)$ and $\text{bin}(n_2, \pi_2)$ mass functions. We can express this in terms of $\delta = \pi_1 - \pi_2$ and a nuisance parameter λ . For instance, if $\lambda = \pi_1 + \pi_2$, we substitute $\pi_1 = (\lambda + \delta)/2$ and $\pi_2 = (\lambda - \delta)/2$. For $\delta = \delta_0$ and a fixed value of λ , we use this binomial product to calculate the probability that the test statistic is at least as large as observed. The P -value is the supremum of such probabilities calculated over all possible values for λ . This provides a family of tests for the various values of δ_0 . The confidence interval for $\pi_1 - \pi_2$ is the set of δ_0 for which this P -value exceeds α . Analogous unconditional intervals apply for the odds ratio and relative risk.

This approach can also be quite conservative. For details regarding various test statistics, see Chan and Zhang (1999) and Santner et al. (2007). To reduce the degree of conservatism, it is better to invert a single two-sided test than to invert two separate one-sided tests (Agresti and Min 2001). For example, with $n_1 = n_2 = 10$ and binomial outcomes $y_1 = 5$ and $y_2 = 1$, the 95% confidence interval for $\pi_1 - \pi_2$ is $(-0.001, 0.700)$ in inverting a two-sided score test and $(-0.020, 0.741)$ in inverting two one-sided tests.

16.7 ALTERNATIVE ESTIMATION THEORY FOR PARAMETRIC MODELS

This text has primarily used the maximum likelihood (ML) approach to inference. This is, by far, the most common approach for categorical data analysis. We've also presented the Bayesian approach. Other frequentist paradigms have been used, however. This section discusses some of them. These methods have similar asymptotic properties as ML, so the large-sample theory presented earlier in this chapter applies also to them.

16.7.1 Weighted Least Squares for Categorical Data

Weighted least squares (WLS) is an extension of ordinary least squares that permits responses to be correlated and to have nonconstant variance. This and related quasi-likelihood methods introduced in Sections 4.7 and 12.3 are sometimes simpler to apply than ML. Familiarity with the WLS method is useful because:

1. WLS computations have a standard form that is simple to apply for a wide variety of models.
2. Algorithms for calculating ML estimates often consist of iterative use of WLS.
3. When the model holds, WLS and ML estimators are asymptotically equivalent, both falling in the class of best asymptotically normal (BAN) estimators.

By (3), for large samples the WLS and ML estimators are approximately normally distributed around the parameter value, and the ratio of their variances converges to 1. Grizzle, Starmer, and Koch (1969) popularized WLS for categorical data analyses. In honor of them, WLS for such analyses is often called the *GSK method*.

For a response variable Y with J categories, consider multinomial samples of sizes n_1, \dots, n_I at I levels of an explanatory variable or combinations of levels of several explanatory variables. Let $\boldsymbol{\pi} = (\boldsymbol{\pi}_1^T, \dots, \boldsymbol{\pi}_I^T)^T$, where

$$\boldsymbol{\pi}_i = (\pi_{1|i}, \pi_{2|i}, \dots, \pi_{J|i})^T \quad \text{with} \quad \sum_j \pi_{j|i} = 1$$

denotes the conditional distribution of Y at level i . Let \mathbf{p} denote corresponding sample proportions, with \mathbf{V} their $IJ \times IJ$ covariance matrix. When the I samples are independent,

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & & \mathbf{0} \\ & \mathbf{V}_2 & \\ & & \ddots \\ \mathbf{0} & & \mathbf{V}_I \end{bmatrix}.$$

From Section 16.1.4, the covariance matrix of $\sqrt{n_i} \mathbf{p}_i$ is

$$n_i \mathbf{V}_i = \begin{bmatrix} \pi_{1|i}(1 - \pi_{1|i}) & -\pi_{1|i}\pi_{2|i} & \cdots & -\pi_{1|i}\pi_{J|i} \\ -\pi_{2|i}\pi_{1|i} & \pi_{2|i}(1 - \pi_{2|i}) & \cdots & -\pi_{2|i}\pi_{J|i} \\ \vdots & \vdots & & \vdots \\ -\pi_{J|i}\pi_{1|i} & -\pi_{J|i}\pi_{2|i} & \cdots & \pi_{J|i}(1 - \pi_{J|i}) \end{bmatrix}.$$

Each set of proportions has $(J - 1)$ linearly independent elements.

Let \mathbf{F} be a vector of $u \leq I(J - 1)$ response functions

$$\mathbf{F}(\boldsymbol{\pi}) = [F_1(\boldsymbol{\pi}), \dots, F_u(\boldsymbol{\pi})]^T.$$

The WLS approach applies to linear models for \mathbf{F} of form

$$(16.29) \quad \mathbf{F}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is a $q \times 1$ vector of parameters and \mathbf{X} is a $u \times q$ model matrix of known constants having rank q . From Section 10.5.1, loglinear and logit response functions are special cases of $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{C} \log(A\boldsymbol{\pi})$ for certain matrices \mathbf{C} and A .

Let $\mathbf{F}(\mathbf{p})$ denote the sample response functions. We assume that \mathbf{F} has continuous second-order partial derivatives in an open region containing $\boldsymbol{\pi}$. This assumption enables the delta method to determine the large-sample normal distribution for $\mathbf{F}(\mathbf{p})$. The asymptotic covariance matrix of $\mathbf{F}(\mathbf{p})$ depends on the $u \times IJ$ matrix

$$\mathbf{Q} = \left[\frac{\partial F_k(\boldsymbol{\pi})}{\partial \pi_{j|i}} \right]$$

for $k = 1, \dots, u$ and all IJ combinations (i, j) . Linear response models have response functions of form $\mathbf{F}(\boldsymbol{\pi}) = A\boldsymbol{\pi}$ for a matrix of known constants A , in which case $\mathbf{Q} = A$. For the generalized loglinear model $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{C} \log(A\boldsymbol{\pi})$ (recall Sections 10.5.1 and 12.1.4), $\mathbf{Q} = \mathbf{C}[\text{Diag}(A\boldsymbol{\pi})]^{-1} A$. By the

multivariate delta method (Section 16.1.5), the asymptotic covariance matrix of $\mathbf{F}(\mathbf{p})$ is

$$\mathbf{V}_F = \mathbf{Q} \mathbf{V} \mathbf{Q}^T.$$

Let $\hat{\mathbf{V}}_F$ denote the sample version of \mathbf{V}_F , substituting sample proportions in \mathbf{Q} and \mathbf{V} . For subsequent formulas, this matrix must be nonsingular.

16.7.2 Inference Using the WLS Approach to Model Fitting

For the general model (16.30), the WLS estimate of β is

$$\mathbf{b} = (\mathbf{X}^T \hat{\mathbf{V}}_F^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{V}}_F^{-1} \mathbf{F}(\mathbf{p}).$$

This is the β value that minimizes the quadratic form

$$[\mathbf{F}(\mathbf{p}) - \mathbf{X}\beta]^T \hat{\mathbf{V}}_F^{-1} [\mathbf{F}(\mathbf{p}) - \mathbf{X}\beta].$$

The ordinary least-squares estimate, for uncorrelated responses with constant variance, results when $\hat{\mathbf{V}}_F$ is a constant multiple of the identity matrix. The WLS estimator has an asymptotic multivariate normal distribution, with estimated covariance matrix

$$\widehat{\text{cov}}(\mathbf{b}) = (\mathbf{X}^T \hat{\mathbf{V}}_F^{-1} \mathbf{X})^{-1}.$$

The normal distribution improves as the sample size increases and $\mathbf{F}(\mathbf{p})$ is more nearly normally distributed.

The estimate \mathbf{b} yields predicted values $\hat{\mathbf{F}} = \mathbf{X}\mathbf{b}$ for the response functions. When the model holds, $\hat{\mathbf{F}}$ is asymptotically better than $\mathbf{F}(\mathbf{p})$ as an estimator of $\mathbf{F}(\boldsymbol{\pi})$ (Section 16.2.2). The estimated covariance matrix of the predicted values is

$$\hat{\mathbf{V}}_{\hat{\mathbf{F}}} = \mathbf{X}(\mathbf{X}^T \hat{\mathbf{V}}_F^{-1} \mathbf{X})^{-1} \mathbf{X}^T.$$

The test of model goodness of fit uses the residual term

$$W = [\mathbf{F}(\mathbf{p}) - \mathbf{X}\mathbf{b}]^T \hat{\mathbf{V}}_F^{-1} [\mathbf{F}(\mathbf{p}) - \mathbf{X}\mathbf{b}] = \mathbf{F}(\mathbf{p})^T \hat{\mathbf{V}}_F^{-1} \mathbf{F}(\mathbf{p}) - \mathbf{b}^T (\mathbf{X}^T \hat{\mathbf{V}}_F^{-1} \mathbf{X}) \mathbf{b},$$

which compares the sample response functions with their model predicted values. Under $H_0: \mathbf{F}(\boldsymbol{\pi}) - \mathbf{X}\beta = \mathbf{0}$ that the model holds, W is asymptotically chi-squared with $\text{df} = u - q$, the difference between the number of response functions and the number of model parameters.

We can more closely check the model fit by studying the residuals, $\mathbf{F}(\mathbf{p}) - \hat{\mathbf{F}}$. They are orthogonal to the fit $\hat{\mathbf{F}}$, so

$$\text{cov}[\mathbf{F}(\mathbf{p})] = \text{cov}\{[\mathbf{F}(\mathbf{p}) - \hat{\mathbf{F}}] + \hat{\mathbf{F}}\} = \text{cov}[\mathbf{F}(\mathbf{p}) - \hat{\mathbf{F}}] + \text{cov}(\hat{\mathbf{F}}).$$

Thus, the estimated covariance matrix of the residuals equals

$$\text{cov}[\mathbf{F}(\mathbf{p})] - \text{cov}(\hat{\mathbf{F}}) = \hat{\mathbf{V}}_F - \hat{\mathbf{V}}_{\hat{\mathbf{F}}} = \hat{\mathbf{V}}_F - \mathbf{X}(\mathbf{X}^T \hat{\mathbf{V}}_F^{-1} \mathbf{X})^{-1} \mathbf{X}^T.$$

Dividing the residuals by their standard errors yields standardized residuals having large-sample standard normal distributions.

Hypotheses about contrasts and other effects of explanatory variables have form $H_0: \mathbf{C}\beta = \mathbf{0}$, where \mathbf{C} is a known $c \times q$ matrix with $c \leq q$, having rank c . The estimator \mathbf{Cb} of $\mathbf{C}\beta$ is asymptotically normal with mean $\mathbf{0}$ under H_0 and with covariance matrix estimated by $\mathbf{C}(\mathbf{X}^T \hat{\mathbf{V}}_F^{-1} \mathbf{X})^{-1} \mathbf{C}^T$. The Wald statistic

$$(16.30) \quad W_C = \mathbf{b}^T \mathbf{C}^T [\mathbf{C}(\mathbf{X}^T \hat{\mathbf{V}}_F^{-1} \mathbf{X}) \mathbf{C}^T]^{-1} \mathbf{C} \mathbf{b}$$

has an approximate chi-squared null distribution with $\text{df} = c$. This statistic also equals the difference between residual chi-squared statistics for the reduced model implied by H_0 and the full model. For the special case $H_0: \beta_i = 0$, $W_C = b_i^2 / \text{var}(b_i)$ has $\text{df} = 1$.

16.7.3 Scope of WLS Versus ML Estimation

The WLS approach requires estimating the multinomial covariance matrix of sample responses at each setting of the explanatory variables. It is inapplicable when explanatory variables are continuous, since there may be only one observation at each such setting. WLS also becomes less appropriate as the number of explanatory variables increases, since few observations may occur at each of the many combinations of settings. By contrast, in principle, continuous explanatory variables or many explanatory settings are not problematic to ML.

When a certain model holds, with large cell expected frequencies ML and WLS give similar results. Both estimators are in the class of best asymptotically normal estimators. However, practical considerations often favor ML estimation. For example, zero cell counts often adversely affect the WLS approach. The sample response functions may then be ill-defined or have a singular estimated covariance matrix.

WLS shares with quasi-likelihood the feature that inferential results depend only on specifying a model for the mean responses and specifying a variance function and covariance structure (here, based on the multinomial). It does not use the likelihood function for the complete distribution. Thus, inference uses Wald methods.

Historically, an advantage of the WLS approach was computational simplicity. This is not relevant now that software is available for ML analyses and for extensions of WLS (e.g., quasi-likelihood methods such as GEE) that do not have some of its disadvantages. Nonetheless, WLS has close connections with more sophisticated methods. Many algorithms for calculating ML estimates (such as the Fisher scoring method of Section 4.6.4 for GLMs) and quasi-likelihood estimates (such as the GEE method) iteratively use WLS.

16.7.4 Minimum Chi-Squared Estimators

Consider estimation of $\boldsymbol{\pi}$ or $\boldsymbol{\theta}$, assuming a model $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$. Let $\hat{\boldsymbol{\theta}}$ denote a generic estimator of $\boldsymbol{\theta}$, for which $\tilde{\boldsymbol{\pi}} = \boldsymbol{\pi}(\hat{\boldsymbol{\theta}})$ estimates $\boldsymbol{\pi}$. The ML estimator $\hat{\boldsymbol{\theta}}$ maximizes the likelihood. It also minimizes the deviance statistic G^2 comparing observed and fitted proportions (Section 16.3.4). Other estimators minimize other measures of distance between $\boldsymbol{\pi}(\boldsymbol{\theta})$ and \mathbf{p} .

The value $\hat{\boldsymbol{\theta}}$ that minimizes the Pearson statistic

$$X^2[\boldsymbol{\pi}(\boldsymbol{\theta}), \mathbf{p}] = n \sum_i \frac{[p_i - \pi_i(\boldsymbol{\theta})]^2}{\pi_i(\boldsymbol{\theta})}$$

is called the *minimum chi-squared estimate*. It is simpler to calculate the estimate that minimizes the *modified chi-squared statistic*

$$(16.31) \quad X_{\text{mod}}^2[\boldsymbol{\pi}(\boldsymbol{\theta}), \mathbf{p}] = n \sum_i \frac{[p_i - \pi_i(\boldsymbol{\theta})]^2}{p_i}$$

that replaces the denominator by the sample proportion. This *minimum modified chi-squared estimate* is the solution for $\boldsymbol{\theta}$ to the equations

$$\sum_i \frac{\pi_i(\boldsymbol{\theta})}{p_i} \left(\frac{\partial \pi_i(\boldsymbol{\theta})}{\partial \theta_j} \right) = 0, \quad j = 1, \dots, q.$$

Neyman (1949) introduced minimum modified chi-squared estimators. He showed that they and minimum chi-squared estimators are best asymptotically normal (BAN) estimators. When the model holds, they are asymptotically equivalent to ML estimators. Under the model, different estimation methods yield nearly identical estimates of parameters when n is large. When the model does not hold, estimates for different methods can be quite different, even when n is large. The estimators converge to values for which the model gives the best approximation to reality, and this approximation is different when best is defined in terms of minimizing G^2 rather than minimizing X^2 or some other measure.

For any n , minimum modified chi-squared estimates are sometimes identical to WLS estimates. The connection refers to an alternative way of specifying a model, using a set of *constraint equations* for $\boldsymbol{\pi}$,

$$\{g_j(\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_N) = 0\}.$$

For instance, for an $I \times J$ table, the $(I-1)(J-1)$ constraint equations

$$\log \pi_{ij} - \log \pi_{i,j+1} - \log \pi_{i+1,j} + \log \pi_{i+1,j+1} = 0$$

specify the model of independence. The number of constraint equations equals the residual df for the model.

Neyman (1949) noted that minimum modified chi-squared estimates result from minimizing

$$\sum_{i=1}^N \frac{(p_i - \pi_i)^2}{p_i} + \sum_{j=1}^{N-q} \lambda_j g_j(\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_N)$$

with respect to $\boldsymbol{\pi}$, where the $\{\lambda_j\}$ are Lagrange multipliers. When the constraint equations are linear in $\boldsymbol{\pi}$, the resulting estimating equations are linear. Then Bhapkar (1966) showed that these estimators are identical to WLS estimators, and (16.32) equals the WLS residual statistic (Section 16.7.1) for testing model fit. Usually, however, constraint equations are nonlinear in $\boldsymbol{\pi}$, such as for the independence model. The WLS estimator is then the minimum modified chi-squared estimator based on a linearized version of the constraints,

$$g_j(\mathbf{p}) + (\boldsymbol{\pi} - \mathbf{p})^T \partial g_j(\boldsymbol{\pi}) / \partial \boldsymbol{\pi} = 0,$$

with differential vector evaluated at \mathbf{p} .

Berkson (1944, 1955, 1980) was a strong advocate of minimum chi-squared methods. For logistic regression, his *minimum logit chi-squared* estimators minimized a weighted sum of squares between sample logits and linear predictions. Mantel (1985) criticized such methods, noting that their consistency requires group sizes to grow large, whereas ML (or conditional ML, when there are many nuisance parameters) is consistent however information goes to the limit (see also Exercise

16.38).

16.7.5 Minimum Discrimination Information

Kullback (1959) formulated estimation by *minimum discrimination information* (MDI). The discrimination information for two probability vectors $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$ is

$$(16.32) \quad I(\boldsymbol{\pi}; \boldsymbol{\gamma}) = \sum_{i=1}^N \pi_i \log(\pi_i / \gamma_i).$$

This directed Kullback–Leibler distance measure between $\boldsymbol{\pi}$ and $\boldsymbol{\gamma}$ is nonnegative, equaling 0 only when $\boldsymbol{\pi} = \boldsymbol{\gamma}$. Gokhale and Kullback (1978) studied MDI estimates that minimize $I(\boldsymbol{\pi}; \boldsymbol{\gamma})$, subject to model constraints, using $\boldsymbol{\gamma} = \mathbf{p}$ for some problems and $\boldsymbol{\gamma}$ with $\gamma_1 = \gamma_2 = \dots = \gamma_N = 1/N$ for others. Good (1963) conducted related work in the area of *maximum entropy*.

In some cases with $\{\gamma_i = 1/N\}$, the MDI estimator is identical to the ML estimator (Simon 1973). With $\boldsymbol{\gamma} = \mathbf{p}$ it is not ML, but it has similar asymptotic properties, being BAN. Then Gokhale and Kullback recommended testing goodness of fit using twice the minimized value of $I(\boldsymbol{\pi}; \mathbf{p})$. This statistic reverses the roles of \mathbf{p} and $\boldsymbol{\pi}$ relative to G^2 , much as X^2_{mod} in (16.32) reverses their roles relative to X^2 . Both statistics fall in the class of power divergence statistics (Cressie and Read 1984, Exercise 1.34) and have similar asymptotic properties. More generally, we could choose any member of the power divergence statistics and define estimates to be the values minimizing it. Under regularity conditions, they are all BAN.

NOTES

Section 16.1: Delta Method

16.1 Delta method generalized: For details of large-sample theory for categorical data, including the delta method, see Bishop et al. (1975, Chap. 14). In applying the delta method to a function g of an asymptotically normal random vector \mathbf{T}_n , suppose that the first-order, ..., $(a - 1)$ st-order differentials of the function are zero at $\boldsymbol{\theta}$, but the a th-order differential is nonzero. A generalization of the delta method implies that $n^{a/2}[g(\mathbf{T}_n) - g(\boldsymbol{\theta})]$ has limiting distribution involving products of order a of components of a normal random vector. When $a = 2$, the limiting distribution is a quadratic form in a multivariate normal vector, which often relates to a chi-squared variable; in the univariate case, it is $\sigma^2[g'(\boldsymbol{\theta})]/2$ times a χ^2_1 variable (Casella and Berger 2001, p. 244).

16.2 Higher-order asymptotics: Higher-order asymptotic methods such as saddlepoint approximations improve on first-order normal approximations. When there are many nuisance parameters, modified profile likelihood functions are useful. See Brazzale et al. (2007), Brazzale and Davison (2008), Davison et al. (2006), Pierce and Peters (1992), and Strawderman and Wells (1998).

16.3 Bootstrap/jackknife: Resampling methods such as the jackknife and the bootstrap are alternative tools for estimating standard errors and obtaining confidence intervals. They can be helpful when use of the delta method is questionable—for instance, for small samples, highly sparse data, or complex sampling designs. For details, see Davison and Hinkley (1997) and Fay (1985).

Section 16.3: Asymptotic Distributions of Residuals and Goodness-of-Fit Statistics

16.4 Asymptotic theory and sparseness: Andersen (1980), Bishop et al. (1975), Cox (1984), Haberman (1974a), and Watson (1959) provided other proofs or considered related cases to the Pearson–Fisher–Cramér–Rao–Birch results. Haberman (1988) showed that large-sample results for X^2 break down with nonstandard asymptotics, such as when the number of cells N grows as $n \rightarrow \infty$ or when different expected frequencies grow at different rates.

16.5 Freedman–Tukey gof: If Y is Poisson, then for large μ the delta method implies \sqrt{Y} is approximately normal with standard deviation 1/2. This motivates the *Freeman–Tukey goodness-of-fit statistic*, $FT = 4 \sum (\sqrt{y_i} - \sqrt{\hat{\mu}_i})^2$. When the model holds, $FT - X^2$ is also $o_p(1)$ as $n \rightarrow \infty$ (Bishop et al. 1975, p. 514).

16.6 Noncentrality: Drost et al. (1989) gave noncentral approximations using other sequences of alternatives than the local and fixed ones [\(16.20\)](#) and [\(16.21\)](#).

Section 16.5: Small-Sample Significance Tests for Contingency Tables

16.7 Exact tests: For exact treatment of $I \times J$ tables, see Mehta and Patel (1983). For ordered categories, see Agresti et al. (1990). For Monte Carlo estimation of exact P -values, see Agresti et al. (1979), Booth and Butler (1999), Diaconis and Sturmfels (1998), Forster et al. (1996), and Patefield (1982). Gail and Mantel (1977) and Good (1976) gave approximate formulas for the number of tables having certain fixed margins. Freidlin and Gastwirth (1999) extended the unconditional approach to tests for trend in $I \times 2$ tables and conditional independence in several 2×2 tables.

16.8 Ancillarity: Suppose that $(\boldsymbol{\theta}, \lambda)$ has minimal sufficient statistic (T, U) , where λ is a nuisance parameter. Cox and Hinkley (1974, p. 35) defined U to be *ancillary* for $\boldsymbol{\theta}$ if its

distribution depends only on λ , and the distribution of T given U depends only on θ . For 2×2 tables with odds ratio θ and $\lambda = (\pi_{1+}, \pi_{+1})$, let $T = n_{11}$ and $U = (n_{1+}, n_{+1})$. Then U is not ancillary, because its distribution depends on θ as well as λ . Using a definition due to Godambe, Bhapkar (1989) referred to the marginals U as *partial ancillary* for θ . This means that the distribution of the data, given U , depends only on θ , and that for fixed θ , the family of distributions of U for various λ is complete. Liang (1984) gave an alternative definition referring to conditional and unconditional inference being equally efficient.

16.9 Randomized tests, fuzzy inference: For discrete data, it is possible to achieve exactly a desired size by using a randomized decision on the boundary of the critical region. To construct a confidence interval that achieves exactly (a priori) a desired coverage probability, we can invert such randomized tests (Stevens 1950). In practice, this approach is not used because of the undesirability of inferential conclusions being determined by a random number, but see Agresti and Gottard (2007) for details. Geyer and Meeden (2005) proposed a related *fuzzy inference* approach consisting of a graphical portrayal of all such possible randomized confidence intervals.

Section 16.7: Alternative Estimation Theory for Multinomial Models

16.10 WLS, MDI: For details about the WLS approach, see Imrey (2011), Imrey et al. (1981), and Koch et al. (1985). For discussion of minimum chi-squared methods, see Neyman (1949), Rao (1963), Bhapkar (1966), and Koch et al. (1985). For more about minimum discrimination information, see Ireland and Kullback (1968ab), Ireland et al. (1969), Ku et al. (1971), and Gokhale and Kullback (1978).

EXERCISES

Applications

16.1 An advertisement by Schering Corp. for the allergy drug Claritin mentioned that in a pediatric randomized clinical trial, symptoms of nervousness were shown by 4 of 188 patients on loratadine (Claritin), 2 of 262 patients taking placebo, and 2 of 170 patients on chloropheniramine. Conduct an analysis of whether nervousness depends on drug.

16.2 Consider a 3×3 table having entries, by row, of $(4,2,0 / 2,2,2 / 0,2,4)$. Conduct an exact test of independence, using X^2 . Assuming ordered rows and columns and using equally spaced scores, conduct an ordinal exact test. Explain why results differ so much.

16.3 Consider exact tests of independence, given the marginals, for the $I \times I$ table having $n_{ii} = 1$ for $i = 1, \dots, I$, and $n_{ij} = 0$ otherwise. Show that **(a)** tests that order tables by their probabilities, X^2 , or G^2 have P -value = 1.0, and **(b)** the one-sided test that orders tables by an ordinal statistic such as r or $C - D$ has P -value = $(1/I!)$.

16.4 For [Table 16.2](#), apply conditional logistic regression to the model discussed in Section 16.6.7.

- a. Obtain an exact P -value for testing no C effect against the alternative of a positive effect. Construct a 95% confidence interval for the conditional CD odds ratio.
- b. Construct the partial tables relating C to D for the combinations of levels of (A, L) . For the sole partial table having data at both C levels, find a 95% exact confidence interval for the odds ratio and find an exact one-sided P -value. Compare to results using the entire data set. Explain why there is no contribution to inference for tables having only a single positive row total or a single positive column total.
- c. Obtain the ordinary ML fit of the logistic regression model. To investigate the sensitivity of the estimated C effect, find the change in the estimate and SE after adding one observation to the data set, a case with no diarrhea when $(C,A,L) = (1,1,1)$.

Theory and Methods

16.5 Explain why:

- a. If $c \neq 0$, cz_n has the same order as z_n ; that is, $o(cz_n)$ is equivalent to $o(z_n)$ and $O(cz_n)$ is equivalent to $O(z_n)$.
- b. $o(y_n)o(z_n) = o(y_nz_n)$, $O(y_n)O(z_n) = O(y_nz_n)$, $o(y_n)O(z_n) = o(y_nz_n)$.

16.6 If X^2 has an asymptotic chi-squared distribution with fixed df as $n \rightarrow \infty$, then explain why $X^2/n = o_p(1)$.

16.7 a. Use Tchebychev's inequality to show that if $E(X_n) = \mu_n$ and $\text{var}(X_n) = \sigma_n^2 < \infty$, then $(X_n - \mu_n) = O_p(\sigma_n)$.

- b.** Suppose that Y_1, \dots, Y_n are iid with $E(Y_i) = \mu$ and $\text{var}(Y_i) = \sigma^2$. For $\bar{y}_n = (\sum_i Y_i)/n$, apply (a) to show that $\bar{y}_n - \mu = O_p(n^{-1/2})$.

16.8 Let Y be a Poisson random variable with mean μ .

- a.** For a constant $c > 0$, show that

$$E[\log(Y + c)] = \log \mu + (c - \frac{1}{2})/\mu + O(\mu^{-2}).$$

[Hint: Note that $\log(Y + c) = \log \mu + \log[1 + (Y + c - \mu)/\mu]$.]

- b.** For independent Poisson cell counts in a 2×2 table, use (a) to argue that the sample log odds ratio after adding $\frac{1}{2}$ to each cell is a sensible estimator for reducing bias in estimating the log odds ratio.

16.9 Let $p = y/n$ for a binomial variate y . Find the asymptotic distribution of the estimator $[p(1 -$

$p)]^{1/2}$ of the standard deviation. What happens when $\pi = 0.5$?

16.10 Suppose T_n has a Poisson distribution with mean $\lambda = n\mu$, for fixed $\mu > 0$. For large n , show that $\log T_n$ is approximately normal with mean $\log(\lambda)$ and variance λ^{-1} . [Hint: By the central limit theorem, T_n/n is approximately $N(\mu, \mu/n)$ for large n .]

16.11 Refer to the previous exercise.

a. If T_n is Poisson, show $\sqrt{T_n}$ has asymptotic variance $1/4$.

b. For a binomial sample proportion p , show the asymptotic variance of $\sin^{-1}(\sqrt{p})$ (with the angle being measured in radians) is $1/4n$. [This transformation and the one in (a) are *variance stabilizing*, producing variates with asymptotic variances that are the same for all values of the parameter. Traditionally, these transformations were employed to make ordinary least squares applicable to count data. See Cochran (1940) for discussion and ML analyses. Rücker et al. (2008) showed that, in reality, this arc sine transformation does not have variance nearly constant when π is near 0 or near 1 and n is not large.]

16.12 For a multinomial $(n, \{\pi_i\})$ distribution, show the correlation between p_i and p_j is $-[\pi_i\pi_j/(1 - \pi_i)(1 - \pi_j)]^{1/2}$. What does this equal when $\pi_i = 1 - \pi_j$ and $\pi_k = 0$ for $k \neq i, j$?

16.13 An animal population has N species, with population proportion π_i of species i . *Simpson's index of ecological diversity* (Simpson 1949) is $I(\boldsymbol{\pi}) = 1 - \sum_i \pi_i^2$. [Rao (1982) surveyed diversity measures.]

a. Two animals are randomly chosen from the population, with replacement. Show that $I(\boldsymbol{\pi})$ is the probability they are of different species.

b. For proportions \mathbf{p} for a random sample, show that the estimated asymptotic standard error of $I(\mathbf{p})$ is

$$2 \left\{ \left[\sum_i p_i^3 - \left(\sum_i p_i^2 \right)^2 \right] / n \right\}^{1/2}.$$

16.14 For independent Poisson random variables $\{Y_i\}$, show that the estimated asymptotic variance of $\sum_i a_i \log(Y_i)$ is $\sum_i a_i^2/y_i$. [This formula applies to ML estimators of parameters for the saturated loglinear model, which are contrasts of $\{\log(y_i)\}$. Formula (16.9) yields the asymptotic covariance structure of such estimators; see Lee (1977).]

16.15 Assuming independent binomial samples, derive the asymptotic standard error of the log relative risk (Section 3.1.3).

16.16 The sample size may need to be large for the ordinal measure $\hat{\gamma}$ to have an approximate normal distribution when $|\gamma|$ is large. The Fisher-type transform $\xi = \frac{1}{2} \log[(1 + \hat{\gamma})/(1 - \hat{\gamma})]$ (Agresti 2010, p. 217; O'Gorman and Woolson 1988) converges more quickly to normality.

a. Show that the asymptotic variance of ξ equals the asymptotic variance of $\hat{\gamma}$ multiplied by $(1 - \hat{\gamma}^2)^{-2}$.

b. Explain how to construct a confidence interval for ζ and use it to obtain one for γ .

c. Show that $\xi = \frac{1}{2} \log(C/D)$. For 2×2 tables, show that this is half the log odds ratio.

16.17 Let $\phi^2(\mathbf{T}) = \sum_i (T_i - \pi_{i0})^2/\pi_{i0}$. Then $\phi^2(\mathbf{p}) = X^2/n$, where X^2 is the Pearson statistic (1.16) for testing $H_0: \pi_i = \pi_{i0}$, $i = 1, \dots, N$, and $n\phi^2(\boldsymbol{\pi})$ is that test's noncentrality when $\boldsymbol{\pi}$ is the true value. Under H_0 , why does the delta method not yield an asymptotic normal distribution for $\phi^2(\mathbf{p})$? (See Note 16.1.)

16.18 In an $I \times J$ contingency table, let θ_{ij} denote local odds ratio (2.10).

a. Show that asymp. $\text{cov}(\sqrt{n} \log \hat{\theta}_{ij}, \sqrt{n} \log \hat{\theta}_{i+1,j}) = -[\pi_{i+1,j}^{-1} + \pi_{i+1,j+1}^{-1}]$.

b. Show that asymp. $\text{cov}(\sqrt{n} \log \hat{\theta}_{ij}, \sqrt{n} \log \hat{\theta}_{i+1,j+1}) = \pi_{i+1,j+1}^{-1}$.

c. When θ_{ij} and θ_{hk} use mutually exclusive sets of cells, show that asymp.

$$\text{cov}(\sqrt{n} \log \hat{\theta}_{ij}, \sqrt{n} \log \hat{\theta}_{hk}) = 0.$$

16.19 Consider the model for a 2×2 table: $\pi_{11} = \theta^2$, $\pi_{12} = \pi_{21} = \theta(1 - \theta)$, $\pi_{22} = (1 - \theta)^2$ (Exercises 3.31 and 11.39).

a. Find A in (16.10) for this model, and use A to obtain the asymptotic variance of $\hat{\theta}$. (As a check, it is simple to find it directly using the inverse of $-E\partial^2 L/\partial\theta^2$, where L is the log likelihood.) For which θ value is the variance maximized? What is the distribution of $\hat{\theta}$ if $\theta = 0$ or $\theta = 1$?

b. Find the asymptotic covariance matrix of $\sqrt{n}\hat{\pi}$.

16.20 Refer to the model for the calf data in Section 1.5.6. Obtain the asymptotic variance of $\hat{\pi}$.

16.21 Cell counts $\{Y_i\}$ are independent Poisson random variables, with $\mu_i = E(Y_i)$. Consider the model

$$\log \mu = X_a \theta_a, \quad \text{where } \mu = (\mu_1, \dots, \mu_N).$$

Using arguments similar to those in Section 14.2, show that the large-sample covariance matrix of $\hat{\theta}_a$ can be estimated by $[X_a^T \text{Diag}(\hat{\mu}) X_a]^{-1}$, where $\hat{\mu}$ is the ML estimator of μ .

16.22 Use the delta method, with derivatives (16.17), to derive the asymptotic covariance matrix in (16.18) for residuals. Show that this matrix is idempotent.

16.23 In some situations, X^2 and G^2 take similar values. Explain the joint influence on this event of (a) whether the model holds, (b) whether the sample size n is large, and (c) whether the number of cells N is large.

16.24 Construct X and θ in multinomial representation (16.22) for the independence model for an $I \times J$ table. By contrast, show X_a for the corresponding Poisson loglinear model (16.24).

16.25 Using (16.13) and (16.23), derive the asymptotic $\text{cov}(\hat{\pi})$ for a multinomial loglinear model.

16.26 Consider the ML estimator $\hat{\pi}_{ij} = p_{i+}p_{+j}$ of π_{ij} for the independence model, when that model does not hold. Show that $E(p_{i+}p_{+j}) = \pi_{i+}\pi_{+j}(n-1)/n + \pi_{ij}/n$. To what does $\hat{\pi}_{ij}$ converge as n increases?

16.27 Let ζ denote a generic measure of association. For K independent multinomial samples of sizes $\{n_k\}$, suppose that $\sqrt{n_k}(\hat{\zeta}_k - \zeta_k) \xrightarrow{d} N(0, \sigma_k^2)$ as $n_k \rightarrow \infty$. A summary measure is

$$\bar{\zeta} = \frac{\sum_k (n_k/\hat{\sigma}_k^2) \hat{\zeta}_k}{\sum_k (n_k/\hat{\sigma}_k^2)}.$$

a. Show that $\sum_k z_k^2 = V + [\bar{\zeta}^2/\hat{\sigma}^2(\bar{\zeta})]$, where

$$z_k = \frac{n_k^{1/2} \hat{\zeta}_k}{\hat{\sigma}_k}, \quad V = \sum_k \frac{n_k(\hat{\zeta}_k - \bar{\zeta})^2}{\hat{\sigma}_k^2}, \quad \hat{\sigma}^2(\bar{\zeta}) = \left(\sum_k \frac{n_k}{\hat{\sigma}_k^2} \right)^{-1}.$$

b. Suppose that $n \rightarrow \infty$ with $n_k/n \rightarrow \rho_k > 0$, $k = 1, \dots, K$. State the asymptotic chi-squared distribution for each component in the partitioning in (a). Indicate the hypothesis that each tests.

16.28 A $2 \times J$ table with fixed row totals consists of two independent multinomial variates. Another $2 \times J$ table, with fixed column totals, consists of J independent binomial variates.

a. For testing that the multinomial distributions have identical parameters, show that the null distribution of $\{n_{ij}\}$ given the sufficient statistics for the common unknown parameters has the multivariate hypergeometric form.

b. For testing that the binomial distributions have identical parameters, show that the null distribution of $\{n_{ij}\}$ given the sufficient statistic for the common unknown parameter is the same as the one derived in (a).

16.29 A Monte Carlo scheme randomly samples M separate $I \times J$ tables having the observed margins to approximate $P_o = P(X^2 \geq X_o^2)$ for an exact test. Let \hat{P} be the sample proportion of the M tables with $X^2 \geq X_o^2$. Show that $P(|\hat{P} - P_o| \leq B) = 1 - \alpha$ requires that $M \approx z_{\alpha/2}^2 P_o (1 - P_o) / B^2$.

16.30 Exercise 1.26 showed LR and score confidence intervals for a binomial sample of size n with $y = 0$.

a. For the Clopper-Pearson approach, show that the upper bound is $1 - (\alpha/2)^{1/n}$, or approximately $-\log(0.025)/n = 3.69/n$ when $\alpha = 0.05$.

b. For the adaptation of the Clopper-Pearson approach using the mid P -value, show that the upper bound is $1 - \alpha^{1/n}$, or approximately $-\log(0.05)/n = 3/n$ when $\alpha = 0.05$.

16.31 For a flip of a coin, let $\pi = P(\text{head})$. An experiment uses $n = 5$ independent flips. Suppose that truly $\pi = 0.50$. Explain why the probability that the 95% Clopper-Pearson confidence interval contains π equals 1.0. [Hint: Is there any possible y for which both one-sided tests of $H_0: \pi = 0.50$ have P -value ≤ 0.025 ?]

16.32 Consider the 95% score confidence interval for the binomial parameter π . When $y = 1$, show that the lower limit is approximately $0.18/n$; in fact, $0 < \pi < 0.18/n$ then falls in an interval only when $y = 0$. Argue that for large n and π just barely below $0.18/n$ or just barely above $1 - 0.18/n$, the coverage probability is about $e^{-0.18} = 0.84$. Hence, even as $n \rightarrow \infty$, this method can have coverage probability much less than 0.95 (Agresti and Coull 1998; Blyth and Still 1983).

16.33 Show that the conditional ML estimate of θ satisfies $n_{11} = E(n_{11})$ for distribution (16.28).

16.34 For the geometric distribution $p(y) = \pi_y(1 - \pi)$, $y = 0, 1, 2, \dots$, show that equating $P(Y \geq y)$ and $P(Y \leq y)$ to $\alpha/2$ yields the confidence interval $[(\alpha/2)^{1/y}, (1 - \alpha/2)^{1/(y+1)}]$. Show that all π between 0 and $1 - \alpha/2$ never fall above a confidence interval, and hence the actual coverage probability exceeds $1 - \alpha/2$ over this region.

16.35 Consider marginal homogeneity for an $I \times I$ table.

a. Letting $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{A}\boldsymbol{\pi}$, explain how (i) $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{0}$, where \mathbf{A} has $I - 1$ rows, and (ii) $\mathbf{F}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta}$, where \mathbf{A} has $2(I - 1)$ rows and $\boldsymbol{\beta}$ has $I - 1$ elements. In part (ii), show \mathbf{A} , $\boldsymbol{\pi}$, \mathbf{X} , $\boldsymbol{\beta}$ when $I = 3$.

b. Explain how to use WLS to test marginal homogeneity. [This is Bhapkar's test (11.15).]

c. Explain why the minimum modified chi-squared estimates are identical to WLS estimates.

16.36 With WLS, show that $[\mathbf{F}(\mathbf{p}) - \mathbf{X}\boldsymbol{\beta}]^T \hat{\mathbf{V}}_{\mathbf{F}}^{-1} [\mathbf{F}(\mathbf{p}) - \mathbf{X}\boldsymbol{\beta}]$ is minimized by $\mathbf{b} = (\mathbf{X}^T \hat{\mathbf{V}}_{\mathbf{F}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{V}}_{\mathbf{F}}^{-1} \mathbf{F}(\mathbf{p})$.

16.37 The response functions $\mathbf{F}(\mathbf{p})$ have asymptotic covariance matrix $\mathbf{V}_{\mathbf{F}}$. Derive the asymptotic covariance matrices of the WLS model parameter estimator \mathbf{b} and the predicted values $\hat{\mathbf{p}} = \mathbf{X}\mathbf{b}$.

16.38 Let y_i be a $\text{bin}(n_i, \pi_i)$ variate for group i , $i = 1, \dots, N$, with $\{y_i\}$ independent. Consider the model that $\pi_1 = \dots = \pi_N$. Denote the common value by π .

a. Show that the ML estimator of π is $\hat{\pi} = (\sum_i y_i) / (\sum_i n_i)$.

b. The minimum chi-squared estimator $\tilde{\pi}$ is the value of π minimizing

$$\sum_{i=1}^N \frac{[(y_i/n_i) - \pi]^2}{\pi} + \sum_{i=1}^N \frac{[(y_i/n_i) - \pi]^2}{1 - \pi}.$$

The second term results from comparing $(1 - y_i/n_i)$ to $(1 - \pi)$, the proportions in the second category. If $n_1 = \dots = n_N = 1$, show that $\tilde{\pi}$ minimizes $N_p(1 - \pi)/\pi + N(1 - p)\pi/(1 - \pi)$. Hence show

$$\tilde{\pi} = p^{1/2} / [p^{1/2} + (1 - p)^{1/2}].$$

Note the considerable bias toward 1/2 in this estimator.

c. As $N \rightarrow \infty$ with all $n_i = 1$, argue that the ML estimator is consistent but the minimum chi-squared estimator is not (Mantel 1985).

d. For $N = 2$ groups, find the minimum modified chi-squared estimator of π . Compare it to the ML estimator.

¹Most analyses in Sections 16.5 and 16.6 can be implemented with StatXact (Cytel Software) and/or R and SAS routines described at the text website.

CHAPTER 17

Historical Tour of Categorical Data Analysis

This book concludes with an informal historical overview of the evolution of methods for categorical data analysis (CDA). We have seen that categorical scales are pervasive in the social sciences and the biomedical sciences. Not surprisingly, the development of GLMs for categorical responses was fostered by statisticians having ties to the social sciences or to the biomedical sciences.

Only in the last quarter of the twentieth century did these models receive the attention given early in the century to models for continuous data. Regression models for continuous variables evolved out of Francis Galton's breakthroughs in the 1880s. The strong influence of R. A. Fisher, G. Udny Yule, and other statisticians on experimentation in agriculture and biological sciences ensured widespread adoption of regression and ANOVA modeling by the mid-twentieth century. On the other hand, despite influential articles around 1900 by Karl Pearson and Yule on association between categorical variables, models for categorical responses received scant attention until the 1960s. Stigler (2002) noted that even simple two-way contingency tables rarely appeared in scientific literature before 1900, and the analyses that were attempted mainly focused on summarizing margins or reducing the data to 2×2 tables.

The beginnings of CDA were often shrouded in controversy. Key figures in the development of statistical science made groundbreaking contributions, but these statisticians were often in heated disagreement with one another.

17.1 PEARSON-YULE ASSOCIATION CONTROVERSY

Much of the early development of methods for CDA took place in England, and it is fitting that we begin our historical tour in London at the beginning of the twentieth century. The year 1900 is an apt starting point, since in that year Karl Pearson introduced his chi-squared statistic (X^2) and G. Udny Yule presented the odds ratio and related measures of association. Before then, most work focused on descriptive aspects for relatively simple measures. For instance, Goodman and Kruskal (1959) noted that the Belgian social statistician Adolphe Quetelet used the relative risk in 1849.

By 1900, Karl Pearson (1857-1936) was already well known in the statistics community. He was head of a statistical laboratory at University College in London. His work the previous decade included developing a large family of probability distributions (called *Pearson curves*, they included important families with skew, such as the gamma), obtaining the product-moment estimate of the correlation coefficient and finding its standard error, and extending Galton's work on linear regression. In fact, Pearson was a true renaissance man, writing on a wide variety of topics that included art, religion, philosophy, law, socialism, women's rights, physics, genetics, eugenics, and evolution. Pearson's motivation for developing the chi-squared test included testing whether outcomes on a roulette wheel in Monte Carlo varied randomly, checking the fit to various data sets of normal distributions and Pearson curves, and testing statistical independence in two-way contingency tables.

Much of the literature on CDA early in the twentieth century consisted of vocal debates about appropriate ways to summarize association. Pearson's approach assumed that continuous bivariate distributions underlie two-way contingency tables (Pearson 1904, 1913). He argued in favor of approximating a measure, such as the correlation, for the underlying continuum. In 1904, Pearson introduced the term *contingency* as a “measure of the total deviation of the classification from independent probability,” and he introduced measures to describe its extent, such as the tetrachoric correlation (Section 2.4.8). The *mean-square contingency* and the *contingency coefficient* are normalizations of X^2 to the (0, 1) scale. The contingency coefficient (Exercise 3.32) for $I \times J$ tables standardized X^2 to approximate an underlying correlation.

George Udny Yule (1871-1951), a British contemporary of Pearson's, took a different approach. Having completed pioneering work developing multiple regression models and multiple and partial correlation coefficients, Yule turned his attention between 1900 and 1912 to association in contingency tables. He believed that many categorical variables, such as (vaccinated, unvaccinated) and (died, survived), are inherently discrete. Yule defined indices directly using cell counts without assuming an underlying continuum. He popularized the odds ratio θ [which Goodman (2000) noted may first have been proposed by a Hungarian statistician, J. Körösy] and a transformation of it to the $[-1, +1]$ scale, $Q = (\theta - 1)/(\theta + 1)$, now called *Yule's Q* (Exercise 2.38). Discussing one of Pearson's measures that assumes underlying normality, Yule argued (1912, p. 612) that “at best the normal coefficient can only be said to give us in cases like these a hypothetical correlation between supposititious variables. The introduction of needless and unverifiable hypotheses does not appear to me a desirable proceeding in scientific work.” Yule (1903) also showed the potential discrepancy between marginal and conditional associations in contingency tables, later studied by E. H. Simpson (1951) and now called *Simpson's paradox*.

In the first quarter of the twentieth century, Karl Pearson was the rarely challenged leader of statistical science in Britain. Pearson's strong personality did not take kindly to criticism, and he reacted negatively to Yule's ideas, arguing that Yule's measures were unsuitable. For instance, Pearson claimed that their values were unstable, since different collapsings of $I \times J$ tables to 2×2 tables could produce quite different values of the measures. Pearson and D. Heron (1913) filled more than 150 pages of *Biometrika*, a journal he co-founded and edited, with a scathing reply to Yule's criticism. In a passage critical also of Yule's well-received book *An Introduction to the*

Theory of Statistics, they stated “If Mr. Yule’s views are accepted, irreparable damage will be done to the growth of modern statistical theory [Yule’s *Q*] has never been and never will be used in any work done under his [Pearson’s] supervision We regret having to draw attention to the manner in which Mr. Yule has gone astray at every stage in his treatment of association, but criticism of his methods has been thrust on us not only by Mr. Yule’s recent attack, but also by the unthinking praise which has been bestowed on a text-book which at many points can only lead statistical students hopelessly astray.” Pearson and Heron attacked Yule’s “half-baked notions” and “specious reasoning” and argued that Yule would have to withdraw his ideas “if he wishes to maintain any reputation as a statistician.”

In retrospect, Pearson and Yule both had valid points. Some classifications, such as most nominal variables, have no apparent underlying continuous distribution. On the other hand, many applications relate naturally to an underlying continuum, and we’ve seen that latent variable models can motivate many standard models and inferences (e.g., Sections 7.1.1 and 8.2.3). Goodman (1981a,b) noted that the ordinal models presented in Sections 10.4.1 and 10.5.2 provide a sort of reconciliation between Yule and Pearson, since Yule’s odds ratio characterizes models that fit well when underlying distributions are approximately normal.

Half a century after the Pearson–Yule controversy, Leo Goodman and William Kruskal surveyed the development of association measures for contingency tables and made many contributions of their own. Their 1979 book reprinted four influential articles of theirs from the *Journal of the American Statistical Association* on this topic. Initial development of many measures occurred in the nineteenth century. Their 1959 article contains the following quote from M. H. Doolittle in 1887, which illustrates the lack of precision in early attempts to quantify the meaning of *association* even in 2×2 tables: “Having given the number of instances respectively in which things are both thus and so, in which they are thus but not so, in which they are so but not thus, and in which they are neither thus nor so, it is required to eliminate the general quantitative relativity inhering in the mere thingness of the things, and to determine the special quantitative relativity subsisting between the thusness and the soness of the things.” Goodman (2000) added to the historical survey and proposed a new measure.

17.2 R. A. FISHER'S CONTRIBUTIONS

Pearson's disagreements with Yule were minor compared with his later ones with Ronald A. Fisher (1890-1962). Using a geometric representation, Fisher (1922) introduced *degrees of freedom* to characterize the family of chi-squared distributions. Fisher claimed that for tests of independence in $I \times J$ tables, X^2 has $df = (I - 1)(J - 1)$. By contrast, Pearson (1900, 1904) had argued that for any application of X^2 , the index that Fisher later identified as df equals the number of cells minus 1, or $IJ - 1$ for two-way tables. Fisher pointed out, however, that estimating hypothesized cell probabilities using estimated row and column probabilities resulted in an additional $(I - 1) + (J - 1)$ constraints on the fitted values, thus affecting the distribution of X^2 .

Not surprisingly, Pearson (1922) reacted critically to Fisher's suggestion that his df formula was incorrect. He stated: "I hold that such a view [Fisher's] is entirely erroneous, and that the writer has done no service to the science of statistics by giving it broad-cast circulation in the pages of the *Journal of the Royal Statistical Society* I trust my critic will pardon me for comparing him with Don Quixote tilting at the windmill; he must either destroy himself, or the whole theory of probable errors, for they are invariably based on using sample values for those of the sampled population unknown to us." Pearson claimed that using row and column sample proportions to estimate unknown probabilities had negligible effect on large-sample distributions, although he had realized (Pearson 1917) that df must be adjusted when the cell counts have linear constraints. Fisher was unable to get his rebuttal published by the Royal Statistical Society, and he ultimately resigned his membership.

Statisticians soon realized that Fisher was correct, but he maintained much bitterness over this and other dealings with Pearson. In the preface to a later volume of his collected works, he remarked that his 1922 article "had to find its way to publication past critics who, in the first place, could not believe that Pearson's work stood in need of correction, and who, if this had to be admitted, were sure that they themselves had corrected it." Writing about Pearson, he stated: "If peevish intolerance of free opinion in others is a sign of senility, it is one which he had developed at an early age." In Fisher (1926), he was able to dig the knife a bit deeper into the Pearson family using 11,688 2×2 tables randomly generated assuming independence by Karl Pearson's son, E. S. Pearson. Fisher showed that the sample mean of X^2 for these tables was 1.00001, much closer to the 1.0 predicted by his formula for $E(X^2)$ of $df = (I - 1)(J - 1) = 1$ than Pearson's $IJ - 1 = 3$. His daughter, Joan Fisher Box (1978), discussed this and other conflicts between Fisher and Pearson. See Stigler (2008) for a discussion of the error in Pearson's argument, and see Fienberg (1980), Hald (1998, pp. 652–663), Plackett (1983), and Stigler (1999, Chap. 19) for more about this controversy.

Fisher's preeminent reputation among statisticians today accrues mainly from his theoretical work (introducing concepts such as sufficiency, information, and optimal properties of ML estimators) and his methodological contributions to the design of experiments and the analysis of variance. Although not so well known for work in CDA, he made other interesting contributions. Moreover, he made good use of the methods in his applied work. For instance, Fisher was also a famed geneticist. In one article, he used Pearson's goodness-of-fit test to check Mendel's theories of natural inheritance and showed that the fit was *too* good (Section 1.5.4).

Fisher realized the limitations of large-sample methods for laboratory work, and he was at the forefront of advocating specialized small-sample methods. Writing about large-sample methods in the preface to the first edition of his classic text *Statistical Methods for Research Workers*, he stated: "[T]he traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a cannon to shoot a sparrow, but it misses the sparrow! The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data." Fisher was among the first to promote the work by W. S. Gosset (pseudonym "Student") on the t distribution. The fifth edition of *Statistical Methods for Research Workers* (1934) introduced Fisher's exact test for 2×2 contingency tables. In his 1935 book *The Design of Experiments*, Fisher described the tea-tasting experiment (Section 3.5.2)

motivated by his experience at an afternoon tea break while employed at Rothamsted Experimental Station.

The mid-1930s finally saw more attention to model building for categorical responses. Chester Bliss (1935), following up a 1933 report on quantal response methods by J. H. Gaddum, popularized the probit model for applications in toxicology with a binary response. Bliss introduced the term *probit* but used the inverse normal cdf with mean 5 (rather than 0, in order to avoid negative values) and standard deviation 1. Stigler (1986, p. 246) and Finney (1971) attributed the first use of inverse normal cdf transformations of proportions to the German physicist Gustav Fechner in his 1860 book *Elemente der Psychophysik*. See Finney (1971) and McCulloch (2000) for other history of the probit method and Chapter 9 of Cramer (2011) for a survey of the early origins of binary regression models.

In the appendix of Bliss (1935), Fisher (1935b) outlined an algorithm for finding ML estimates of model parameters. That algorithm was a Newton–Raphson type of method using expected information, today commonly called *Fisher scoring* (Section 4.6.2). Fisher (1954) argued for using ML for binary models with an appropriate link function in place of the popular approach at the time of applying a variance-stabilizing transformation in order to use ordinary least-squares methods.

The definition for homogeneous association (no interaction) in contingency tables originated in an article by the British statistician Maurice Bartlett (1935) about $2 \times 2 \times 2$ tables. Bartlett showed how to solve a cubic equation to find ML estimates of cell probabilities satisfying the property of equality of odds ratios between two variables at each level of the third. He attributed the idea to Fisher. In the same year, Sam Wilks proposed the likelihood-ratio test of independence in contingency tables.

In 1940, Fisher developed canonical correlation methods for contingency tables. He showed how to assign scores to rows and columns of a contingency table to maximize the correlation. His work relates to the later development, particularly in France, of *correspondence analysis* methods (e.g., Benzécri 1973). In the same year, Deming and Stephan showed how to apply iterative proportional fitting (IPF) for raking contingency tables to maintain associations while satisfying fixed marginal distributions.

R. A. Fisher has been the greatest influence on the practice of modern statistical science. The biography by his daughter (Box 1978) gives a fascinating account of his impressive contributions to statistics and genetics. Fienberg (1980) summarized his contributions to CDA.

17.3 LOGISTIC REGRESSION

The logit transform showed up sporadically before its use in binomial logistic regression. For example, Bartlett (1937) used $\log[y/(1 - y)]$ in regression and ANOVA to transform observations y that are continuous proportions (Exercise 4.35), and in a book of statistical tables published in 1938, R. A. Fisher and Frank Yates suggested it as a possible transformation of a binomial parameter for analyzing binary data. In 1944, the physician and statistician Joseph Berkson introduced the term *logit* for this transformation, and following Wilson and Worcester (1943), who had employed it for estimating LD₅₀, proposed the logistic regression model. Berkson showed that the logistic model fitted similarly to the probit model, and his subsequent work did much to popularize the model. He argued, however, for fitting using the computationally simpler minimum logit chi-squared rather than ML. [See also his comments following Fisher (1954) in this regard.] In 1951, Jerome Cornfield, another statistician with strong medical ties, used the odds ratio to approximate relative risks in case-control studies. Dyke and Patterson (1952) apparently first used the logit in models with qualitative predictors.

Sir David Cox introduced many statisticians to logistic regression, through his influential 1958 article and 1970 book, *The Analysis of Binary Data*. About the same time, an article by the Danish statistician and mathematician Georg Rasch sparked an enormous literature on item-response models. The most important of these is the logit model with subject and item parameters, now called the *Rasch model* (Section 13.1.4). This work was highly influential in the psychometric community of northern Europe (especially in Denmark, the Netherlands, and Germany) and spurred many generalizations in the educational testing community in the United States.

The extension of logistic regression to multicategory responses received occasional attention before 1970 (e.g., Mantel 1966) but substantial work after about that date. For nominal responses, early work was mainly in the econometrics literature. See Bock (1970), McFadden (1974), and Theil (1969, 1970). In 2000, Daniel McFadden won the Prize in Economic Sciences in Memory of Alfred Nobel for his work in the 1970s and 1980s on the discrete-choice model (Section 8.5). For cumulative logit models for ordinal responses, see Bock and Jones (1968), Simon (1974), Snell (1964), Walker and Duncan (1967), and Williams and Grizzle (1972). The cumulative probit case, shown to result from a normal latent variable model (McKelvey and Zavoina 1975), has a longer history; see, for instance, Aitchison and Silvey (1957) and Bock and Jones (1968, Chap. 8). Cumulative logit and probit models received much more attention following publication of McCullagh (1980), which provided a Fisher scoring algorithm for ML fitting of all cumulative link models.

The next major advances with logistic regression dealt with its application to case-control studies (e.g., Breslow 1996, Mantel 1973, Prentice 1976a, Prentice and Pyke 1979; see also Section 5.1.4) and the conditional ML approach to model fitting for those studies and others with numerous nuisance parameters (Breslow et al. 1978, with related work cited in Note 7.5). The conditional approach was later exploited in small-sample exact inference (Hirji et al. 1987, Mehta and Patel 1995). See also Sections 7.3, 11.2, 16.5, and 16.6.

Nathan Mantel, whose name appears in the preceding two paragraphs, made a variety of interesting contributions to CDA. Although best known for the 1959 Mantel-Haenszel test and related odds ratio estimator, he also discussed trend tests (1963), multinomial logit and loglinear modeling (1966), logistic regression for case-control data (1973), the number of contingency tables having fixed margins (Gail and Mantel 1977), the analysis of square contingency tables (Mantel and Byar 1978), and problems with minimum chi-squared and Wald tests (1985, 1987a).

Logistic regression has become a useful component of causal inference methods and methods of dealing with missing data. An example is the introduction by Rosenbaum and Rubin (1983) of the *propensity score* for modeling the probability of being in some treatment group, as a device of adjusting for bias in treatment comparisons with observational studies.

More recently, attention has focused on fitting logistic models to correlated responses for clustered

data. One strand of this is marginal modeling of longitudinal data (Liang and Zeger 1986, Liang et al. 1992, Lipsitz et al. 1994). Much of this literature focuses on quasi-likelihood methods such as generalized estimating equations (GEEs). Another strand is generalized linear mixed models (e.g., Breslow and Clayton 1993, Pierce and Sands 1975).

Perhaps the most far-reaching contribution of the past half century has been the introduction by British statisticians John Nelder and R. W. M. Wedderburn in 1972 of the concept of *generalized linear models*. This unifies the logistic and probit regression models for binomial data with loglinear models for Poisson data and with long-established regression and ANOVA models for normal-response data. Interestingly, the algorithm they used to fit GLMs is Fisher scoring, which R. A. Fisher introduced in 1935 for ML fitting of probit models. McCulloch (2000) reviewed the journey from probit models to GLMs and their further generalizations such as quasi-likelihood.

17.4 MULTIWAY CONTINGENCY TABLES AND LOGLINEAR MODELS

The quarter century following the end of World War II saw the development of a theoretical underpinning for models for contingency tables. H. Cramér (1946) and C. R. Rao (1957, 1963) derived general expressions for large-sample distributions of parameter estimators.

In 1949, the Berkeley-based statistician Jerzy Neyman, who had already performed fundamental work on hypothesis testing and interval estimation methods with E. S. Pearson, introduced the family of *best asymptotically normal* (BAN) estimators. These have the same optimal large-sample properties as ML estimators. The BAN family includes estimators obtained by minimizing chi-squared-type measures comparing observed proportions to proportions predicted by the model (Section 16.7.4). This type of estimator itself includes some *weighted least squares* (WLS) estimators. The simplicity of their computation, compared with ML estimators, was an important consideration before the advent of modern computing. Neyman's (1949) only mention of Fisher was the suggestion that Fisher did not realize that estimators other than ML could be BAN, stating that "the results ... contradict the assertion of R. A. Fisher, not a very clear one, that 'the maximum likelihood equation may indeed be derived from the conditions that it shall be linear in frequencies, and efficient for all values of θ '." In fact, Fisher had realized in the 1920s that other estimators could be efficient (see Stigler's "The epic story of maximum likelihood" 2007 article in *Statist. Sci.*), and he often returned the jab at Neyman, such as in writing (1956) about proposals for an unconditional test for 2×2 tables, "the principles of Neyman and Pearson's 'Theory of Testing Hypotheses' are liable to mislead those who follow them into much wasted effort."

In the early 1950s, William Cochran published work dealing with a variety of important topics in CDA. Scottish-born, Cochran spent most of his career at American universities: Iowa State, North Carolina State, Johns Hopkins, and Harvard. Cochran (1940) modeled Poisson and binomial responses with variance-stabilizing transformations. He (1943) recognized and discussed ways of dealing with overdispersion. His comments following Fisher (1954) also dealt with practical issues, such as dangers of relying solely on ML methods when extra heterogeneity existed beyond what standard distributions assume. Cochran (1950) introduced a generalization (Cochran's Q) of McNemar's test for comparing proportions in several matched samples. His classic 1954 article is a mixture of new methodology and advice for applied statisticians. It gave sample-size guidelines for chi-squared approximations to work well for the X^2 statistic, pointing out that the guideline that expected frequencies should exceed 5 was often too strict. It also stressed the importance of directing inferences toward narrow (e.g., single-degree-of-freedom) alternatives and partitioning chi-squared statistics into components. One instance of this was Cochran's proposed test of conditional independence in several 2×2 tables, which was closely related to the Mantel and Haenszel (1959) test (Section 6.4.2). Another was a test for a linear trend in proportions across quantitatively defined rows of an $I \times 2$ table (Section 5.3.5). See also Cochran (1955). Fienberg (1984) reviewed Cochran's contributions to CDA.

Bartlett's work on interaction structure in $2 \times 2 \times 2$ contingency tables had relatively little impact for 20 years. Indeed, in presenting methods for partitioning X^2 in $2 \times 2 \times 2$ tables, Lancaster (1951) noted that "Doubtless little use will ever be made of more than a three-dimensional classification." However, in the mid-1950s and early 1960s, Bartlett's work was extended in many ways to multiway tables. See, for instance, Darroch (1962), Good (1963), Goodman (1964b), Plackett (1962), Roy and Kastenbaum (1956), and Roy and Mitra (1956). These articles as well as influential articles by Martin W. Birch (1963, 1964a,b, 1965) were the genesis of research work on loglinear models between about 1965 and 1975. Birch's work was part of a never-submitted Ph.D. thesis at the University of Glasgow. He explicitly provided loglinear model formulas and explained analogies with factorial ANOVA models, and he showed how to obtain ML estimates of cell probabilities in three-way tables for various models. He showed the equivalence of those ML estimates for Poisson and multinomial sampling. He (and Watson 1959) extended theoretical results of Cramér and Rao on

large-sample distributions for contingency table models.

A survey article by the French statistician Henri Caussinus (1966), based partly on his Ph.D. thesis, provides a good glimpse of the state-of-the-art of CDA in the middle of these two decades of advances. There, Caussinus introduced the quasi-symmetry model for square tables. Issue number 4 in Volume XI of *Annales de la Faculté des Sciences de Toulouse Mathématiques*, a special 2002 issue honoring Caussinus at his retirement, contains remembrances by Caussinus about the origins of this contribution as well as several articles investigating this property and its links and extensions.

Much of the work in the next decades on loglinear and related logit modeling took place at three American universities: the University of Chicago, Harvard University, and the University of North Carolina. At Chicago, Leo Goodman wrote a series of groundbreaking articles, dealing with such topics as partitionings of chi-squared, models for square tables (e.g., quasi-independence), stepwise logit and loglinear model-building procedures, deriving asymptotic variances of ML estimates of loglinear parameters, latent class models (building on early work by Paul Lazarsfeld), association models, correlation models, and correspondence analysis. For surveys of his early work, see Goodman (1968, an R. A. Fisher memorial lecture, 1970). For later work, see Goodman (1985, 1996, 2000). Goodman also wrote a stream of articles for social science journals that had a substantial impact on popularizing loglinear and logit methods for applications (e.g., Goodman 2007 and references therein). (See [Figure 17.1](#).)

Figure 17.1 Four leading figures in the development of categorical data analysis.



Karl Pearson



G. Udny Yule



Ronald A. Fisher



Leo Goodman

Over the past 60 years, Goodman has been the most prolific contributor to the advancement of CDA methodology. The field owes tremendous gratitude to his steady and impressive body of work. In addition, some of Goodman's students at Chicago also made fundamental contributions. In 1970,

Shelby Haberman completed a Ph.D. dissertation (the basis of his 1974a monograph) making substantial theoretical contributions to loglinear modeling. Among topics he considered were residual analyses, existence of ML estimates, loglinear models for ordinal variables, and theoretical results for models (such as the Rasch model) for which the number of parameters grows with the sample size. Clifford Clogg followed in Goodman's footsteps by having influence in the social sciences and in statistics with his work on association models, demography, models for rates, the census, and various other topics.

Simultaneously with Goodman's work, related research on ML methods for loglinear–logit models occurred at Harvard by students of Frederick Mosteller (such as Stephen Fienberg) and William Cochran. Much of this research was inspired by problems arising in analyzing large, multivariate data sets in the National Halothane Study (see Chap. 5 in Mosteller's 2010 autobiography, *The Pleasures of Statistics*). That study investigated whether halothane was more likely than other anesthetics to cause death due to liver damage. A presidential address by Mosteller (1968) to the American Statistical Association described early uses of loglinear models for smoothing multidimensional discrete data sets. Yvonne Bishop (1969) noted the equivalence between loglinear and logit models and showed the usefulness of IPF for model fitting. Fienberg (1970ab) dealt with theoretical aspects of IPF as well as existence of ML estimates for square-table models. For the past 40 years, he has been one of the most prolific researchers in loglinear modeling, together with many of his Ph.D. students. A landmark book in 1975 by Bishop and Fienberg with Paul Holland, *Discrete Multivariate Analysis*, was largely responsible for introducing loglinear models to the general statistical community and remains an important reference.

Research at North Carolina by Gary Koch and colleagues and many of his Ph.D. students (such as J. Richard Landis, Peter Imrey, and Maura Stokes) has been highly influential in the biomedical sciences. Their research developed WLS methods for categorical data models (Section 16.7.1). The 1969 article by Koch with his colleagues J. Grizzle and F. Starmer popularized this approach. Koch and colleagues extended it in later articles to an impressive variety of problems, including problems for which ML methods are awkward to use, such as the analysis of repeated categorical measurement data (Koch et al. 1977). In 1966, Vasant Bhapkar showed that the WLS estimator is often identical to Neyman's minimum modified chi-squared estimator. Imrey (2011) surveyed Koch's contributions, and Fienberg (2011) related the UNC work to related developments elsewhere. (See [Figure 17.2](#).)

Figure 17.2 Alan Agresti with statisticians from the North Carolina (Peter Imrey, Gary Koch, J. Richard Landis) and Harvard (Stephen Fienberg) schools of CDA. Photo taken by Bahjat Qaqish at 2009 UNC Festschrift for Gary Koch.



The early literature on loglinear models treated all classifications as nominal. Haberman (1974b) and Simon (1974) showed how to exploit ordinality of classifications in loglinear models. This work was extended in several articles by Leo Goodman (1979a, 1981a,b, 1983). The extensions included association models, which can replace ordered scores in loglinear models by parameters (Section 10.5). Goodman (1985, 1986, 1996) also discussed related correlation models and provided a model-based perspective for essentially equivalent correspondence analysis methods. Joseph Lang

(1996a, 2004, 2005) has extended ML fitting to broad classes of models, including generalized loglinear models.

Certain loglinear models with conditional independence structure provide graphical models for contingency tables (Section 10.1.2). The article by Darroch et al. (1980) was the genesis of much of this work.

Fienberg and Rinaldo (2007) provided a historical overview of the development of loglinear models. That article also discusses issues still to be resolved adequately, such as whether ML estimates exist for large, sparse contingency tables containing many sampling zeroes.

17.5 BAYESIAN METHODS FOR CATEGORICAL DATA

We now summarize the development of Bayesian methods for categorical data analysis. This actually dates back 250 years, as Bayes in 1763 (and then Laplace in 1774) estimated a binomial parameter using a uniform prior distribution. See Stigler (1986, pp. 100–136) for details.

Early applications of Bayesian methods to contingency tables involved smoothing cell counts to improve estimation of cell probabilities. In particular, large sparse tables often contain many sampling zeros, for which 0.0 is unappealing as a probability estimate. Also, Stein’s results for estimating multivariate normal means suggest that lower total mean squared error occurs with Bayes estimators that shrink the sample proportions toward some average value (Efron and Morris 1975).

I. J. Good (1956) used log-normal and gamma priors in estimating *association factors* in contingency tables (Section 2.4.2). Good’s (1965) monograph summarized the use of Bayesian methods for estimating multinomial probabilities in contingency tables, using a Dirichlet prior distribution. Good (1967) focused on suitable priors for multinomial probabilities in significance tests and made considerable efforts to reconcile Bayesian and frequentist inference, such as by relating Bayes factors to chi-squared statistics. He was innovative in his early use of hierarchical and empirical Bayesian approaches. Much of Good’s early work was apparently motivated by his collaborations with Alan Turing at Bletchley Park, England, during World War II in work toward breaking the German code for its wartime communications. Albert (2010) reviewed Good’s early research as well as later related work on smoothing contingency tables. By contrast, early critics of the Bayesian approach included R. A. Fisher (1956), who challenged the use of a uniform prior for the binomial parameter, noting that uniform priors on other scales would lead to different results.

For 2×2 tables, Altham (1969) gave a Bayesian analysis comparing parameters for two independent binomial samples, using independent beta priors (Section 3.6.2). Seneta and Phipps (2001) noted that a Swiss medical doctor, Carl Liebermeister, had suggested such an approach with uniform priors in 1877. Altham (1971) showed Bayesian analyses for binomial proportions from matched-pairs data. The Bayesian approaches presented by then focused directly on cell probabilities by using a prior distribution for them. In an influential article, Lindley (1964) focused on estimating summary measures of association in contingency tables. For instance, using a Dirichlet prior distribution for the multinomial probabilities, he found the posterior distribution of contrasts of log probabilities, such as the log odds ratio. An alternative approach (Leonard 1975, Laird 1978) focused on parameters of the saturated loglinear model, using normal priors. The approach of using normal priors for logits received considerable attention in the 1970s by Leonard and others (e.g., Leonard 1972).

In the context of model selection for analyzing contingency tables, Raftery (1986) proposed replacing P -values by Bayes factors. He suggested BIC as a simple approximation to $2[\log(\text{Bayes factor})]$. Since then, there has been an enormous literature on issues of model selection including model averaging (e.g., Madigan and Raftery 1994). BIC itself has become an increasingly popular alternative to AIC, but for criticisms of it, see articles by Gelman and Rubin and others in the February 1999 issue of *Sociological Methods and Research*. Spiegelhalter et al. (2002) proposed a deviance information criterion (DIC) as a hierarchical modeling generalization of the AIC.

The difficulty of calculating the posterior distribution when the prior is not conjugate is less problematic with modern ways of approximating posterior distributions by simulating samples from them. These include the importance sampling generalization of Monte Carlo simulation (Zellner and Rossi 1984) and Markov chain Monte Carlo methods such as Gibbs sampling (Gelfand and Smith 1990). Zellner and Rossi used Bayesian methods with importance sampling for logistic regression and Gelfand and Smith considered a class of multinomial models with Dirichlet prior. Zeger and Karim (1991) fitted generalized linear mixed models (GLMMs) essentially using a Bayesian framework with priors for fixed and random effects.

The Bayesian literature on CDA methodology has exploded in the past 25 years since the

introduction of MCMC methods. For further details and references, see Agresti and Hitchcock (2005, also at the text website), Congdon (2005), Leonard (1999), and Leonard and Hsu (1994).

17.6 A LOOK FORWARD, AND BACKWARD

Methods for categorical data analysis have developed in dramatic fashion over the past half century. In many ways, the area is now a relatively mature one, and it seems unlikely that the development will be nearly as dramatic in the next half century. However, it is unwise to think that this can be predicted without considerable uncertainty.

As in all branches of statistics, it does seem safe to predict that in coming years a primary topic for development will be methods for dealing with data sets with very large numbers of variables. With modeling methods, there is the challenge of developing adequate model checking and diagnostic methods. In the Bayesian context, there is the challenge of adequate specification of prior distributions with huge numbers of parameters so that those priors are not overly influential in the analysis. Some research on methods for large numbers of variables is largely outside the realm of traditional modeling, such as the *data mining* methods briefly introduced in Chapter 15. For these and other complex data structures and applications that place a premium on predictive power, methodologists will need to find ways to overcome the sacrifice of simplicity and interpretability of structure. Important areas of application are likely to continue to include genetics, such as the analysis of discrete DNA sequences in the form of very high-dimensional contingency tables, and business applications such as credit scoring and market basket analysis for predicting future behavior of customers.

As sources for the historical tour in this chapter, I would like to especially acknowledge Stigler (1986), *Studies in the History of Probability and Statistics*, edited by E. S. Pearson and M. G. Kendall (London: Griffin, 1970), and personal conversations over the years with many statisticians, including Erling Andersen, R. L. Anderson, Henri Caussinus, Herman Chernoff, William Cochran, Sir David Cox, John Darroch, Leo Goodman, David Hoaglin, Gary Koch, Frederick Mosteller, John Nelder, Ingram Olkin, C. R. Rao, Donald Rubin, Stephen Stigler, Geoffrey Watson, and Marvin Zelen.

To readers who have made it this far, I congratulate your perseverance! To gain a more complete view of the historical development of CDA, you may want to read articles such as Fienberg and Rinaldo (2007), Goodman (2000), and Imrey et al. (1981, 1996), or browse through some early books on this topic, such as R. L. Plackett's *The Analysis of Categorical Data* (London: Griffin, 1974) and the Bishop, Fienberg, and Holland *Discrete Multivariate Analysis* (Cambridge, MA: MTT Press 1975). Finally, you may want to browse the following chronological list of 28 sources, which convey a sense of how methodology has evolved.

- Pearson (1900)
- Yule (1912)
- Fisher (1922)
- Bartlett(1935)
- Berkson (1944)
- Neyman (1949)
- Cochran (1954)
- Goodman and Kruskal (1954)
- Roy and Mitra (1956)
- Cox (1958a)
- Mantel and Haenszel (1959)
- Birch (1963)
- Caussinus(1966)
- Goodman (1968)
- Mosteller(1968)
- Grizzle et al. (1969)
- Goodman (1970)

Haberman (1974a)
Nelder and Wedderburn (1972)
McFadden (1974)
Goodman (1979a)
McCullagh (1980)
Liang and Zeger (1986)
Breslow and Clayton (1993)
Albert and Chib (1993)
Bickel and Levina (2004)
Lang (2004)
Hastie, Tibshirani, and Friedman (2009)

APPENDIX A

Statistical Software for Categorical Data Analysis

In this appendix we very briefly summarize statistical software for categorical data analysis. A much more detailed appendix, *Using Statistical Software for Categorical Data Analysis*, is at the text website:

www.stat.ufl.edu/~aa/cda/cda.html

That appendix presents details about software use for all of the methods presented in this text, with separate sections for R, SAS, SPSS, and Stata. It also shows code for R and SAS for many examples in this text. We have placed it there rather than in the hard copy of the book itself (1) because of the rather long length of this edition, (2) so it can be updated easily over time as software capabilities change, and (3) to make it easier to copy particular examples as you conduct your own analyses.

A.1 SAS

SAS is general-purpose software for a wide variety of statistical analyses. The main procedures (PROCs) for categorical data analyses are FREQ, GENMOD, LOGISTIC, NLMIXED, GLIMMIX, and CATMOD.

PROC FREQ computes confidence limits for the binomial proportion including score, Jeffreys Bayes, Agresti–Coull, and Clopper–Pearson intervals, equivalence and noninferiority tests for the binomial proportion and the proportion difference, unconditional exact confidence limits for the proportion (risk) difference, measures of association and their estimated standard errors, multinomial goodness-of-fit tests, tests of independence in $I \times J$ tables including exact small-sample methods, generalized Cochran–Mantel–Haenszel tests of conditional independence, and the Zelen exact test for equal odds ratios.

PROC GENMOD fits generalized linear models using ML or Bayesian methods. It also fits cumulative link models for ordinal responses and zero-inflated Poisson regression models for count data. It can perform GEE analyses for marginal models and gives ML fitting of binary response models, cumulative link models for ordinal responses, and baseline–category logit models for nominal responses. It also can perform conditional logistic regression and small-sample inference using the conditional likelihood. PROC CATMOD fits baseline-category logit models. It is also useful for weighted least-squares fitting of a wide variety of models for nonsparse contingency tables. PROC SURVEY-LOGISTIC can fit binary and multiple-category logistic models by the method of pseudo maximum likelihood, incorporating the sample design into the analysis.

PROC NLMIXED fits generalized linear mixed models (GLMMs). It approximates the likelihood using adaptive Gauss-Hermite quadrature. PROC GLIMMIX also fits such models with a variety of fitting methods, including pseudo likelihood methods, and provides built-in distributions and associated variance functions as well as link functions for categorical responses.

Other programs run on SAS that are not specifically supported by the SAS Institute. For further details about SAS for categorical data analyses, see the very helpful guide by Stokes et al. (2012).

A.2 R AND S-PLUS

R is free open-source software maintained and regularly updated by many volunteers. See www.r-project.org, at which site you can download it and find various documentation.

Dr. Laura Thompson has prepared an excellent, detailed manual on the use of S-PLUS and R to conduct the analyses shown in the second edition of this book. There is a link to it at the text website (www.stat.ufl.edu/~aa/cda/cda.html). There are also links there to statisticians who have online material using R for categorical data analyses. A useful book on statistical modeling using R is by Aitkin et al. (2009).

Some useful R functions for categorical data analysis are:

- *prop.test* for a test and score CI for a binomial proportion
- *chisq.test* for the chi-squared test of independence
- *fisher.test* for Fisher's exact test
- *glm* for generalized linear models such as logistic regression, Poisson regression, and loglinear models

R can do various other analyses using specialized functions available in libraries or from certain people. Examples are:

- Functions for forming score and other confidence intervals for proportions and measures such as the difference of proportions and odds ratio, at the text website, www.stat.ufl.edu/~aa/cda/cda.html.
- A function *mph.fit* written by Prof. Joseph Lang (joseph-lang@uiowa.edu) for ML fitting of the generalized loglinear model (10.10), marginal models, and the much more general "multinomial-Poisson homogeneous" models considered in Lang (2004, 2005).
- The VGAM library (www.stat.auckland.ac.nz/~yee/VGAM) and its *vglm* function written by Thomas Yee, which can fit a wide variety of models for multinomial response variables and other types of discrete data.

A.3 STATA

For examples of categorical data analyses for many data sets in my text *An Introduction to Categorical Data Analysis*, see the useful site www.ats.ucla.edu/stat/examples/icda set up by the UCLA Statistical Computing Center. In Stata, the programs:

- *tabulate* can generate many measures of association and their standard errors
- *glm* can fit generalized linear models such as logistic regression and loglinear models
- *mlogit* can fit baseline-category logit models and *ologit* can fit ordinal models
- the *GLLAMM* module (www.gllamm.org) can fit a very wide variety of models, including logistic and cumulative logit models with random effects

A.4 SPSS

On the *Analyze* menu, the *Descriptive statistics* option has a *Crosstabs* suboption that provides several methods for contingency tables, including measures of association and their standard errors. The *Generalized linear models* option has a *Generalized linear models* suboption, the *Regression* option has a *Binary logistic* suboption and an *Ordinal* suboption for a cumulative link model and *Multinomial logistic* suboption for a baseline-category logit model, the *Loglinear* option has a *General* suboption, and the *Generalized linear models* option has the *Generalized estimating equations (EE)* suboption. For further details on all of the above, see the text website.

A.5 STATXACT AND LOGXACT

The Cytel Software package *StatXact* (www.cytel.com/Software/statXact) provides small-sample confidence intervals for a binomial parameter, the difference of proportions, relative risk, and odds ratio. It provides Fisher's exact test and its generalizations for $I \times J$ tables and can conduct exact tests of conditional independence in stratified tables and tests of equality of odds ratios, and can construct exact confidence intervals for the common odds ratio in several 2×2 tables. Its companion *LogXact* (www.cytel.com/Software/Logxact) performs exact conditional logistic regression for categorical responses. Their manuals are good resources for summaries of small-sample methods.

A.6 OTHER SOFTWARE

The text website also provides information about other software, such as for HLM and MLwiN for multilevel models, LATENT GOLD and LEM for latent class models, LIMDEP and NLOGIT for multinomial discrete-choice models, SUDAAN for survey data, and SUPERMIX for generalized linear mixed models.

APPENDIX B

Chi-Squared Distribution Values

df	Right-Tailed Probability						
	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	7.84	10.64	12.59	14.45	16.81	18.55	22.46
7	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	10.22	13.36	15.51	17.53	20.09	21.96	26.12
9	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	13.70	17.28	19.68	21.92	24.72	26.76	31.26
12	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	23.83	28.41	31.41	34.17	37.57	40.00	45.32
25	29.34	34.38	37.65	40.65	44.31	46.93	52.62
30	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	45.62	51.80	55.76	59.34	63.69	66.77	73.40
50	56.33	63.17	67.50	71.42	76.15	79.49	86.66
60	66.98	74.40	79.08	83.30	88.38	91.95	99.61
70	77.58	85.53	90.53	95.02	100.4	104.2	112.3
80	88.13	96.58	101.8	106.6	112.3	116.3	124.8
90	98.65	107.6	113.1	118.1	124.1	128.3	137.2
100	109.1	118.5	124.3	129.6	135.8	140.2	149.5

References

- Agresti, A. 1992. A survey of exact inference for contingency tables. *Stat. Sci.* **7**: 131–153.
- Agresti, A. 1993. Computing conditional maximum likelihood estimates for generalized Rasch models using simple loglinear models with diagonal parameters. *Scand. J. Stat.* **20**: 63–71.
- Agresti, A. 1997. A model for repeated measurements of a multivariate binary response. *J. Am. Stat. Assoc.* **92**: 315–321.
- Agresti, A. 1999. On logit confidence intervals for the odds ratio with small samples. *Biometrics* **55**: 597–602.
- Agresti, A. 2001. Exact inference for categorical data: Recent advances and continuing controversies. *Stat. Med.* **20**: 2709–2722.
- Agresti, A. 2010. *Analysis of Ordinal Categorical Data*, 2nd ed. Hoboken, NJ: Wiley.
- Agresti, A. 2011. Score and pseudo-score confidence intervals for categorical data analysis. *Stat. Biopharm. Res.* **3**: 163–172.
- Agresti, A., and B. Caffo. 2000. Simple and effective confidence intervals for proportions and difference of proportions result from adding two successes and two failures. *Am. Stat.* **54**: 280–288.
- Agresti, A., and B. A. Coull. 1998. Approximate is better than exact for interval estimation of binomial parameters. *Am. Stat.* **52**: 119–126.
- Agresti, A., and A. Gottard. 2007. Nonconservative exact small-sample inference for discrete data. *Comput. Stat. Data An.* **51**: 6447–6458.
- Agresti, A., and J. Hartzel. 2000. Strategies for comparing treatments on a binary response with multi-centre data. *Stat. Med.* **19**(8): 1115–1139.
- Agresti, A., and D. Hitchcock. 2005. Bayesian inference for categorical data analysis. *Stat. Methods Applic.* **14**: 297–330.
- Agresti, A., and J. Lang. 1993. A proportional odds model with subject-specific effects for repeated ordered categorical responses. *Biometrika* **80**: 527–534.
- Agresti, A., and I. Liu. 1999. Modeling a categorical variable allowing arbitrarily many category choices. *Biometrics* **55**: 936–943.
- Agresti, A., and Y. Min. 2001. On small-sample confidence intervals for parameters in discrete distributions. *Biometrics* **57**: 963–971.
- Agresti, A., and Y. Min. 2003. Effects and non-effects of paired identical observations in comparing proportions with binary matched-pairs data. *Stat. Med.* **23**: 65–75.
- Agresti, A., and Y. Min. 2005a. Frequentist performance of Bayesian confidence intervals for comparing proportions in 2×2 contingency tables. *Biometrics* **61**: 515–523.
- Agresti, A., and Y. Min. 2005b. Improved confidence intervals for comparing matched proportions. *Stat. Med.* **24**: 729–740.
- Agresti, A., and R. Natarajan. 2001. Modeling clustered ordered categorical data: A survey. *Int. Stat. Rev.* **69**: 345–371.
- Agresti, A., and E. Ryu. 2010. Pseudo-score confidence intervals for parameters in discrete statistical models. *Biometrika* **97**: 215–222.
- Agresti, A., D. Wackerly, and J. Boyett. 1979. Exact conditional tests for cross-classifications: Approximation of attained significance levels. *Psychometrika* **44**: 75–84.
- Agresti, A., C. Chuang, and A. Kezouh. 1987. Order-restricted score parameters in association models for contingency tables. *J. Am. Stat. Assoc.* **82**: 619–623.
- Agresti, A., C. R. Mehta, and N. R. Patel. 1990. Exact inference for contingency tables with ordered categories. *J. Am. Stat. Assoc.* **85**: 453–458.

- Agresti, A., J. Booth, J. Hobert, and B. Caffo. 2000. Random-effects modeling of categorical response data. *Sociol. Methodol.* **30**: 27–81.
- Agresti, A., B. Caffo, and P. Ohman-Strickland. 2004. Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. *Comput. Stat. Data An.* **47**: 639–653.
- Agresti, A., M. Bini, B. Bertaccini, and E. Ryu. 2008. Simultaneous confidence intervals for comparing binomial parameters. *Biometrics* **64**: 1270–1275.
- Aitchison, J., and C. G. G. Aitken. 1976. Multivariate binary discrimination by the kernel method. *Biometrika* **63**: 413–420.
- Aitchison, J., and J. A. Bennett. 1970. Polychotomous quantal response by maximum indicant. *Biometrika* **57**: 253–262.
- Aitchison, J., and S. M. Shen. 1980. Logistic-normal distributions: Some properties and uses. *Biometrika* **67**: 261–272.
- Aitchison, J., and S. D. Silvey. 1957. The generalization of probit analysis to the case of multiple responses. *Biometrika* **44**: 131–140.
- Aitchison, J., and S. D. Silvey. 1958. Maximum likelihood estimation of parameters subject to restraints. *Ann. Math. Stat.* **29**: 813–828.
- Aitkin, M. 1979. A simultaneous test procedure for contingency table models. *Appl. Stat.* **28**: 233–242.
- Aitkin, M. 1999. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**: 117–128.
- Aitkin, M., and D. Clayton. 1980. The fitting of exponential, Weibull, and extreme value distributions to complex censored survival data using GLIM. *Appl. Stat.* **29**: 156–163.
- Aitkin, M., D. Anderson, and J. Hinde. 1981. Statistical modelling of data on teaching styles. *J. R. Stat. Soc. Ser. A* **144**: 419–461.
- Aitkin, M., B. J. Francis, J. P. Hinde, and R. E. Darnell. 2009. *Statistical Modelling in R*. Oxford: Oxford University Press.
- Albert, A., and J. A. Anderson. 1984. On the existence of maximum likelihood estimates in logistic models. *Biometrika* **71**: 1–10.
- Albert, J. 1997. Bayesian testing and estimation of association in a two-way contingency table. *J. Am. Stat. Assoc.* **92**: 685–693.
- Albert, J. 2010. Good smoothing. In *Frontiers of Statistical Decision Making and Bayesian Analysis*, eds. M.-H. Chen, D. K. Dey, P. Müller, D. Sun, and K. Ye. New York: Springer, pp. 419–436.
- Albert, J., and S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.* **88**: 669–679.
- Allison, P. D. 1982. Discrete-time methods for the analysis of event histories. *Sociol. Methodol.* **13**: 61–98.
- Allison, P. D. 1999. Comparing logit and probit coefficients across groups. *Sociol. Methods Res.* **28**: 185–208.
- Altham, P. M. E. 1969. Exact Bayesian analysis of a 2×2 contingency table and Fisher's "exact" significance test. *J. R. Stat. Soc. Ser B* **31**: 261–269.
- Altham, P. M. E. 1970. The measurement of association of rows and columns for an $r \times s$ contingency table. *J. R. Stat. Soc. Ser B* **32**: 63–73.
- Altham, P. M. E. 1971. The analysis of matched proportions. *Biometrika* **58**: 561–576.
- Altham, P. M. E. 1975. Quasi-independent triangular contingency tables. *Biometrics* **31**: 233–238.
- Altham, P. M. E. 1978. Two generalizations of the binomial distribution. *Appl. Stat.* **27**: 162–167.

- Altham, P. M. E. 1984. Improving the precision of estimation by fitting a model. *J. R. Stat. Soc. Ser B* **46**: 118–119.
- Altham, P. M. E. 2010. Using recently developed software on a 2×2 table of matched pairs with incompletely classified data. *Appl. Stat.* **59**: 377–379.
- Amemiya, T. 1981. Qualitative response models: A survey. *J. Econom. Literature* **19**: 1483–1536.
- Andersen, A. H. 1974. Multidimensional contingency tables. *Scand. J. Stat.* **1**: 115–127.
- Andersen, E. B. 1970. Asymptotic properties of conditional maximum-likelihood estimators. *J. R. Stat. Soc. Ser B* **32**: 283–301.
- Andersen, E. B. 1980. *Discrete Statistical Models with Social Science Applications*. Amsterdam: North-Holland.
- Andersen, E. B. 1995. Polytomous Rasch models and their estimation. In *Rasch Models: Foundations, Recent Developments, and Applications*, eds. G. Fischer and I. Molenaar. New York: Springer-Verlag, pp. 272–291.
- Anderson, C. J., and U. Böckenholt. 2000. Graphical regression models for polytomous variables. *Psychometrika* **65**: 497–509.
- Anderson, C. J., and J. K. Vermunt. 2000. Log-multiplicative models as latent variable models for nominal and/or ordinal data. *Sociol. Methodol.* **30**: 81–121.
- Anderson, C. J., and J.-T. Yu. 2007. Log-multiplicative association models as item response models. *Psychometrika* **72**: 5–23.
- Anderson, C. J., S. Wasserman, and B. Crouch. 1999. A p* primer: logit models for social networks. *Social Networks* **21**: 37–66.
- Anderson, D. A., and M. Aitkin. 1985. Variance component models with binary response: Interviewer variability. *J. R. Stat. Soc. Ser B* **47**: 203–210.
- Anderson, J. A. 1972. Separate sample logistic discrimination. *Biometrika* **59**: 19–35.
- Anderson, J. A. 1975. Quadratic logistic discrimination. *Biometrika* **62**: 149–154.
- Anderson, J. A. 1984. Regression and ordered categorical variables. *J. R. Stat. Soc. Ser B* **46**: 1–30.
- Anderson, J. A., and P. R. Philips. 1981. Regression, discrimination, and measurement models for ordered categorical variables. *Appl. Stat.* **30**: 22–31.
- Anderson, T. W., and L. A. Goodman. 1957. Statistical inference about Markov chains. *Ann. Math. Stat.* **28**: 89–110.
- Aranda-Ordaz, F.J. 1981. On two families of transformations to additivity for binary response data. *Biometrics* **68**: 357–363.
- Aranda-Ordaz, F.J. 1983. An extension of the proportional hazards model for grouped data. *Biometrics* **39**: 109–117.
- Armitage, P. 1955. Tests for linear trends in proportions and frequencies. *Biometrics* **11**: 375–386.
- Ashford, J. R., and R. D. Sowden. 1970. Multivariate probit analysis. *Biometrics* **26**: 535–546.
- Asmussen, S., and D. Edwards. 1983. Collapsibility and response variables in contingency tables. *Biometrika* **70**: 567–578.
- Azzalini, A. 1994. Logistic regression for autocorrelated data with application to repeated measures. *Biometrika* **81**: 767–775.
- Azzalini, A., A. Bowman, and W. Härdle. 1989. On the use of nonparametric regression for model checking. *Biometrika* **76**: 1–11.
- Azzalini, A., and B. Scarpa. 2012. *Data Analysis and Data Mining*. Oxford: Oxford University Press.
- Baglivo, J., D. Olivier, and M. Pagano. 1992. Methods for exact goodness-of-fit tests. *J. Am. Stat. Assoc.* **87**: 464–469.
- Baker, F. B., and L. J. Hubert. 1975. Measuring the power of hierarchical cluster analysis. *J. Am.*

Stat. Assoc. **70**: 31–38.

- Baker, S. G. 1994. The multinomial–Poisson transformation. *The Statistician* **43**: 495–504.
- Baker, S. G., and N. M. Laird. 1988. Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *J. Am. Stat. Assoc.* **83**: 62–69.
- Banerjee, C., M. Capozzoli, L. McSweeney, and D. Sinha. 1999. Beyond kappa: A review of interrater agreement measures. *Can. J. Stat.* **27**: 3–23.
- Banerjee, S., B. P. Carlin, and A. E. Gelfand. 2004. *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: CRC Press.
- Baptista, J., and M. C. Pike. 1977. Algorithm AS115: Exact two-sided confidence limits for the odds ratio in a 2×2 table. *Appl. Stat.* **26**: 214–220.
- Barnard, G. A. 1945. A new test for 2×2 tables. *Nature* **156**: 177.
- Barnard, G. A. 1947. Significance tests for 2×2 tables. *Biometrika* **34**: 123–138.
- Barnard, G. A. 1949. Statistical inference. *J. R. Stat. Soc. Ser B* **11**: 115–139.
- Barnard, G. A. 1979. In contradiction to J. Berkson's dispraise: Conditional tests can be more efficient. *J. Stat. Plan. Infer.* **3**: 181–188.
- Bartholomew, D. J. 1980. Factor analysis for categorical data. *J. R. Stat. Soc. B* **42**: 293–321.
- Bartholomew, D. J., M. Knott, and I. Moustaki. 2011. *Latent Variable Models and Factor Analysis: A Unified Approach*, 3rd ed. Hoboken, NJ: Wiley.
- Bartlett, M. S. 1935. Contingency table interactions. *J. R. Stat. Soc. Suppl.* **2**: 248–252.
- Bartlett, M. S. 1937. Some examples of statistical methods of research in agriculture and applied biology. *J. R. Stat. Soc. Suppl.* **4**: 137–183.
- Bartolucci, F., and A. Forcina. 2002. Extended RC association models allowing for order restrictions and marginal modeling. *J. Am. Stat. Assoc.* **97**: 1192–1199.
- Becker, M. 1989a. Models for the analysis of association in multivariate contingency tables. *J. Am. Stat. Assoc.* **84**: 1014–1019.
- Becker, M. 1989b. On the bivariate normal distribution and association models for ordinal categorical data. *Stat. Probab. Lett.* **8**: 435–440.
- Becker, M., and A. Agresti. 1992. Log-linear modelling of pairwise interobserver agreement on a categorical scale. *Stat. Med.* **11**: 101–114.
- Becker, M., and C. C. Clogg. 1989. Analysis of sets of two-way contingency tables using association models. *J. Am. Stat. Assoc.* **84**: 142–151.
- Beder, J. H., and R. C. Heim. 1990. On the use of ridit analysis. *Psychometrika* **55**: 603–616.
- Bedrick, E. J. 1983. Chi-squared tests for cross-classified tables of survey data. *Biometrika* **70**: 591–595.
- Bedrick, E. J., R. Christensen, and W. Johnson. 1997. Bayesian binomial regression: Predicting survival at a trauma center. *Am. Stat.* **51**: 211–218.
- Begg, C. B. 1990. On inferences from Wei's biased coin design for clinical trials. *Biometrika* **77**: 467–484.
- Begg, C. B., and R. Gray. 1984. Calculation of polytomous logistic regression parameters using individualized regressions. *Biometrika* **71**: 11–18.
- Beggs, S., S. Cardell, and J. Hausman. 1981. Assessing the potential demand for electric cars. *J. Econometrics* **16**: 1–19.
- Beitler, P. J., and J. R. Landis. 1985. A mixed-effects model for categorical data. *Biometrics* **41**: 991–1000.
- Bell, R. M., Y. Koren, and C. Volinsky. 2010. All together now: A perspective on the Netflix prize. *Chance* **23**(1): 24–29.
- Benedetti, J. K., and M. B. Brown. 1978. Strategies for the selection of loglinear models. *Biometrics*

- Benichou, J. 2005. Attributable risk. In *Encyclopedia of Biostatistics*, 2nd ed. Chichester, UK: Wiley, pp. 249–262.
- Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**: 289–300.
- Benjamini, Y., and Yekutieli, D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**: 1165–1188.
- Benzécri, J.-P. 1973. *L'Analyse des Données*, Vol. 1, *La Taxonomie*; Vol. 2, *L'Analyse des Correspondances*. Paris: Dunod.
- Berger, R., and D. D. Boos. 1994. *p*-Values maximized over a confidence set for the nuisance parameter. *J. Am. Stat. Assoc.* **89**: 1012–1016.
- Bergsma, W. P., and T. Rudas. 2002. Marginal models for categorical data. *Ann. Stat.* **30**: 140–159.
- Bergsma, W. P., M. A. Croon, and J. A. Hagenaars. 2009. *Marginal Models: For Dependent, Clustered, and Longitudinal Categorical Data*. New York: Springer.
- Berkson, J. 1938. Some difficulties of interpretation encountered in the application of the chi-square test. *J. Am. Stat. Assoc.* **33**: 526–536.
- Berkson, J. 1944. Application of the logistic function to bio-assay. *J. Am. Stat. Assoc.* **39**: 357–365.
- Berkson, J. 1951. Why I prefer logits to probits. *Biometrics* **7**: 327–339.
- Berkson, J. 1953. A statistically precise and relatively simple method of estimating the bioassay with quantal response, based on the logistic function. *J. Am. Stat. Assoc.* **48**: 565–599.
- Berkson, J. 1955. Maximum likelihood and minimum logit χ^2 estimation of the logistic function. *J. Am. Stat. Assoc.* **50**: 130–162.
- Berkson, J. 1978. In dispraise of the exact test. *J. Stat. Plan. Infer.* **2**: 27–42.
- Berkson, J. 1980. Minimum chi-square, not maximum likelihood! *Ann. Stat.* **8**: 457–487.
- Berry, G., and P. Armitage. 1995. Mid-*P* confidence intervals: A brief review. *The Statistician* **44**: 417–423.
- Besag, J. E. 1974. Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. B* **36**: 192–236.
- Bhadra, D., M. Daniels, S. Kim, M. Ghosh, and B. Mukherjee. 2012. A Bayesian semiparametric approach for incorporating longitudinal information on exposure history for inference in case-control studies. *Biometrics* **68**: 361–370.
- Bhapkar, V. P. 1966. A note on the equivalence of two test criteria for hypotheses in categorical data. *J. Am. Stat. Assoc.* **61**: 228–235.
- Bhapkar, V. P. 1968. On the analysis of contingency tables with a quantitative response. *Biometrics* **24**: 329–338.
- Bhapkar, V. P. 1973. On the comparison of proportions in matched samples. *Sankhyā Ser. A* **35**: 341–356.
- Bhapkar, V. P. 1989. Conditioning on ancillary statistics and loss of information in the presence of nuisance parameters. *J. Stat. Plan. Infer.* **21**: 139–160.
- Bhapkar, V. P., and J. N. Darroch. 1990. Marginal symmetry and quasi symmetry of general order. *J. Multiv. Anal.* **34**: 173–184.
- Bhapkar, V. P., and G. W. Somes. 1977. Distribution of Q when testing equality of matched proportions. *J. Am. Stat. Assoc.* **72**: 658–661.
- Bickel, P. J., and E. Levina. 2004. Some theory for Fisher's linear discriminant function, 'naive Bayes,' and some alternatives when there are many more variables than observations. *Bernoulli* **10**: 989–1010.
- Billingsley, P. 1961. Statistical methods in Markov chains. *Ann. Math. Stat.* **32**: 12–40.

- Birch, M. W. 1963. Maximum likelihood in three-way contingency tables. *J. R. Stat. Soc. B* **25**: 220–233.
- Birch, M. W. 1964a. A new proof of the Pearson–Fisher theorem. *Ann. Math. Stat.* **35**: 817–824.
- Birch, M. W. 1964b. The detection of partial association I: The 2×2 case. *J. R. Stat. Soc. B* **26**: 313–324.
- Birch, M. W. 1965. The detection of partial association II: The general case. *J. R. Stat. Soc. B* **27**: 111–124.
- Bishop, Y. M. M. 1969. Full contingency tables, logits, and split contingency tables. *Biometrics* **25**: 383–399.
- Bishop, Y. M. M. 1971. Effects of collapsing multidimensional contingency tables. *Biometrics* **27**: 545–562.
- Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland. 1975. *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press.
- Blaker, H. 2000. Confidence curves and improved exact confidence intervals for discrete distributions. *Can. J. Stat.* **28**: 783–798.
- Blanchard, G., O. Bousquet, and P. Massart. 2008. Statistical performance of support vector machines. *Ann. Stat.* **36**: 489–531.
- Bliss, C. I. 1935. The calculation of the dosage–mortality curve. *Ann. Appl. Biol.* **22**: 134–167.
- Bloch, D. A., and G. S. Watson. 1967. A Bayesian study of the multinomial distribution. *Ann. Math. Stat.* **38**: 1423–1435.
- Blyth, C. R. 1972. On Simpson’s paradox and the sure-thing principle. *J. Am. Stat. Assoc.* **67**: 364–366.
- Blyth, C. R. 1980. Expected absolute error of the usual estimator of the binomial parameter. *Am. Stat.* **34**: 155–157.
- Blyth, C. R., and H. A. Still. 1983. Binomial confidence intervals. *J. Am. Stat. Assoc.* **78**: 108–116.
- Bock, R. D. 1970. Estimating multinomial response relations. In *Contributions to Statistics and Probability*, ed. R. C. Bose. Chapel Hill, NC: University of North Carolina Press, pp. 453–479.
- Bock, R. D., and M. Aitkin. 1981. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **46**: 443–459.
- Bock, R. D., and L. V. Jones. 1968. *The Measurement and Prediction of Judgement and Choice*. San Francisco: Holden-Day.
- Böckenholt, U., and W. Dillon. 1997. Modelling within-subject dependencies in ordinal paired comparison data. *Psychometrika* **62**: 411–434.
- Bonney, G. E. 1987. Logistic regression for dependent binary observations. *Biometrics* **43**: 951–973.
- Boos, D. D. 1992. On generalized score tests. *Am. Stat.* **46**: 327–333.
- Booth, J., and R. Butler. 1999. An importance sampling algorithm for exact conditional tests in log-linear models. *Biometrika* **86**: 321–332.
- Booth, J. G., and J. P. Hobert. 1998. Standard errors of prediction in generalized linear mixed models. *J. Am. Stat. Assoc.* **93**: 262–272.
- Booth, J. G., and J. P. Hobert. 1999. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Stat. Soc. B* **61**: 265–285.
- Booth, J. G., G. Casella, H. Friedl, and J. P. Hobert. 2003. Negative binomial loglinear mixed models. *Stat. Modelling* **3**: 179–191.
- Booth, J. G., G. Casella, and J. P. Hobert. 2008. Clustering using objective functions and stochastic search. *J. R. Stat. Soc. B* **70**: 119–139.
- Boschloo, R. D. 1970. Raised conditional level of significance for the 2×2 table when testing the

- equality of two probabilities. *Stat. Neerland.* **24**: 1–9.
- Bowker, A. H. 1948. A test for symmetry in contingency tables. *J. Am. Stat. Assoc.* **43**: 572–574.
- Box, J. F. 1978. *R. A. Fisher: The Life of a Scientist*. Hoboken, NJ: Wiley.
- Bradley, R. A. 1976. Science, statistics, and paired comparisons. *Biometrics* **32**: 213–240.
- Bradley, R. A., and M. E. Terry. 1952. Rank analysis of incomplete block designs I. The method of paired comparisons. *Biometrika* **39**: 324–345.
- Brazzale, A. R. and A. C. Davison. 2008. Accurate parametric inference for small samples. *Stat. Sci.* **23**: 465–484.
- Brazzale, A. R., A. C. Davison, and N. Reid. 2007. *Applied Asymptotics: Case Studies in Small-Sample Statistics*. Cambridge, UK: Cambridge University Press.
- Breiman L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Boca Raton, FL: CRC Press.
- Breslow, N. 1976. Regression analysis of the log odds ratio: A method for retrospective studies. *Biometrics* **32**: 409–416.
- Breslow, N. 1981. Odds ratio estimators when the data are sparse. *Biometrika* **68**: 73–84.
- Breslow, N. 1984. Extra-Poisson variation in log-linear models. *Appl. Stat.* **33**: 38–44.
- Breslow, N. 1990. Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *J. Am. Stat. Assoc.* **85**: 565–571.
- Breslow, N. 1996. Statistics in epidemiology: The case-control study. *J. Am. Stat. Assoc.* **91**: 14–28.
- Breslow, N., and D. G. Clayton. 1993. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* **88**: 9–25.
- Breslow, N., and N. E. Day. 1980, 1987. *Statistical Methods in Cancer Research*, Vol. I, *The Analysis of Case-Control Studies*; Vol. II. *The Design and Analysis of Cohort Studies*. Lyon: IARC.
- Breslow, N., and X. Lin. 1995. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82**: 81–91.
- Breslow, N., and W. Powers. 1978. Are there two logistic regressions for retrospective studies? *Biometrics* **34**: 100–105.
- Breslow, N., N. Day, K. Halvorsen, R. Prentice, and C. Sabai. 1978. Estimation of multiple relative risk functions in matched case-control studies. *Am. J. Epidemiol.* **108**: 299–307.
- Brier, S. S. 1980. Analysis of contingency tables under cluster sampling. *Biometrika* **67**: 591–596.
- Brooks, S. P., B. J. T. Morgan, M. S. Ridout, and S. E. Pack. 1997. Finite mixture models for proportions. *Biometrics* **53**: 1097–1115.
- Bross, I. D. J. 1958. How to use ridit analysis. *Biometrics* **14**: 18–38.
- Brass, I. D. J. 1967. Pertinency of an extraneous variable. *J. Chronic Diseases* **20**: 487–497.
- Brown, L., and X. Li. 2005. Confidence intervals for two sample binomial distribution. *J. Stat. Plan. Infer.* **130**: 359–375.
- Brown, M. B. 1976. Screening effects in multidimensional contingency tables. *Appl. Stat.* **25**: 37–46.
- Brown, M. B., and J. K. Benedetti. 1977. Sampling behavior of tests for correlation in two-way contingency tables. *J. Am. Stat. Assoc.* **72**: 309–315.
- Brown, P. J., and P. W. K. Rundell. 1985. Kernel estimates for categorical data. *Technometrics* **27**: 293–299.
- Brown, L. D., T. T. Cai, and A. DasGupta. 2001. Interval estimation for a binomial proportion. *Stat. Sci.* **16**: 101–133.
- Browne, W. J., S. V. Subramanian, K. Jones, and H. Goldstein. 2005. Variance partitioning in multilevel logistic models that exhibit overdispersion. *J. R. Stat. Soc. A* **168**: 599–613.
- Brusco, M. J. 2004. Clustering binary data in the presence of masking variables. *Psych. Methods* **9**:

- Bull, S. B., and A. Donner. 1987. The efficiency of multinomial logistic regression compared with multiple group discriminant analysis. *J. Am. Stat. Assoc.* **82**: 1118–1122.
- Buonaccorsi, J. 2010. *Measurement Error: Models, Methods, and Applications*. Boca Raton, FL: CRC Press.
- Burnham, K. P., and D. Anderson. 2010. *Model Selection and Multi-Model Inference*, 2nd ed. New York: Springer.
- Burnham, K. P. and W. S. Overton. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* **65**: 625–633.
- Burr, D., and H. Doss. 2005. A Bayesian semiparametric model for random-effects meta-analysis. *J. Am. Stat. Assoc.* **100**: 242–251.
- Burridge, J. 1981. A note on maximum likelihood estimation for regression models using grouped data. *J. R. Stat. Soc. B* **43**: 41–45.
- Caffo, B., and M. Griswold. 2006. A user-friendly introduction to link-probit-normal models. *Am. Stat.* **60**: 139–145.
- Caffo, B., M. W. An, and C. Rohde. 2007. Flexible random intercept models for binary outcomes using mixtures of normals. *Comput. Stat. Data An.* **51**: 5220–5235.
- Cameron, A. C., and P. K. Trivedi. 1998. *Regression Analysis of Count Data*. Cambridge, UK: Cambridge University Press.
- Campbell, I. 2007. Chi squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations. *Stat. Med.* **26**: 3661–3675.
- Carey, V., S. L. Zeger, and P. Diggle. 1993. Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80**: 517–526.
- Carlin, J. B., R. Wolfe, C. H. Brown, and A. Gelman. 2001. A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. *Biostatistics* **2**: 397–416.
- Carroll, R. J., S. Wang, and C. Y. Wang. 1995. Prospective analysis of logistic case–control pairs. *J. Am. Stat. Assoc.* **90**: 157–169.
- Casella, G., and R. Berger. 2001. *Statistical Inference*, 2nd ed. Pacific Grove, CA: Wadsworth.
- Casella, G., and E. Moreno. 2005. Intrinsic meta-analysis of contingency tables. *Stat. Med.* **28**: 583–604.
- Casella, G., and E. Moreno. 2009. Assessing robustness of intrinsic tests of independence in two-way contingency tables. *J. Am. Stat. Assoc.* **104**: 1261–1271.
- Catalano, P. J., and L. M. Ryan. 1992. Bivariate latent variable models for clustered discrete and continuous outcomes. *J. Am. Stat. Assoc.* **87**: 651–658.
- Caussinus, H. 1966. Contribution à l’analyse statistique des tableaux de corrélation. *Ann. Fac. Sci. Univ. Toulouse* **29**: 77–182.
- Cerioli, A. 2002. Testing mutual independence between two discrete-values spatial processes: A correction to Pearson chi-squared. *Biometrics* **58**: 888–897.
- Chafaï, D. 2009. Confidence regions for the multinomial parameter with small sample size. *J. Am. Stat. Assoc.* **104**: 1071–1079.
- Chaganty, N. R., and H. Joe. 2004. Efficiency of generalized estimating equations for binary responses. *J. R. Stat. Soc. B* **66**: 851–860.
- Chaganty, N. R., and H. Joe. 2006. Range of correlation matrices for dependent Bernoulli random variables. *Biometrika* **93**: 197–206.
- Chaloner, K., and K. Larntz. 1989. Optimal Bayesian design applied to logistic regression experiments. *J. Stat. Plan. Infer.* **21**: 191–208.

- Chamberlain, G. 1980. Analysis of covariance with qualitative data. *Rev. Econ. Stud.* **47**: 225–238.
- Chambers, E. A., and D. R. Cox. 1967. Discrimination between alternative binary response models. *Biometrika* **54**: 573–578.
- Chambers, R. L., and D. G. Steel. 2001. Simple methods for ecological inference in 2×2 tables. *J. R. Stat. Soc. A* **164**: 175–192.
- Chan, I. 1998. Exact tests of equivalence and efficacy with non-zero lower bound for comparative studies. *Stat. Med.* **17**: 1403–1413.
- Chan, I. S. F., and Z. Zhang. 1999. Test-based exact confidence intervals for the difference of proportions. *Biometrics* **55**: 1202–1209.
- Chao, A., P. K. Tsay, S.-H. Lin, W.-Y. Shau, and D.-Y. Chao. 2001. The applications of capture–recapture models to epidemiological data. *Stat. Med.* **20**: 3123–3157.
- Chen, M.-H., and Q.-M. Shao. 1999. Properties of prior and posterior distributions for multivariate categorical response data models. *J. Multiv. Anal.* **71**: 277–296.
- Chen, M. H., J. G. Ibrahim, and S. Kim. 2008. Properties and implementation of Jeffreys's prior in binomial regression models. *J. Am. Stat. Assoc.* **103**: 1659–1664.
- Chen, Z. and L. Kuo. 2001. A note on the estimation of the multinomial logit model with random effects. *Am. Stat.* **55**: 89–95.
- Cheng, P. E., M. Liou, and J. A. D. Aston. 2010. Likelihood ratio tests with three-way tables. *J. Am. Stat. Assoc.* **105**: 740–749.
- Cheng, P. E., M. Liou, J. A. D. Aston, and A. C. Tsai. 2008. Information identities and testing hypotheses: Power analysis for contingency tables. *Stat. Sin.* **18**: 535–558.
- Chib, S., and E. Greenberg. 1998. Analysis of multivariate probit models. *Biometrika* **85**: 347–361.
- Choulakian, V. 1988. Exploratory analysis of contingency tables by loglinear formulation and generalizations of correspondence analysis. *Psychometrika* **53**: 235–250.
- Christensen, R., W. Johnson, A. Branscum, and T. E. Hanson. 2010. *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. Boca Raton, FL: CRC Press.
- Clogg, C. C. 1995. Latent class models. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, eds. G. Arminger and C. C. Clogg. New York: Plenum Press, pp. 311–359.
- Clogg, C. C., and L. A. Goodman. 1984. Latent structure analysis of a set of multidimensional contingency tables. *J. Am. Stat. Assoc.* **79**: 762–771.
- Clogg, C. C., and E. S. Shihadeh. 1994. *Statistical Models for Ordinal Variables*. Thousand Oaks, CA: Sage Publications.
- Clopper, C. J., and E. S. Pearson. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**: 404–413.
- Cochran, W. G. 1940. The analysis of variance when experimental errors follow the Poisson or binomial laws. *Ann. Math. Stat.* **11**: 335–347.
- Cochran, W. G. 1943. Analysis of variance for percentages based on unequal numbers. *J. Am. Stat. Assoc.* **38**: 287–301.
- Cochran, W. G. 1950. The comparison of percentages in matched samples. *Biometrika* **37**: 256–266.
- Cochran, W. G. 1952. The χ^2 test of goodness-of-fit. *Ann. Math. Stat.* **23**: 315–345.
- Cochran, W. G. 1954. Some methods of strengthening the common χ^2 tests. *Biometrics* **10**: 417–451.
- Cochran, W. G. 1955. A test of a linear function of the deviations between observed and expected numbers. *J. Am. Stat. Assoc.* **50**: 377–397.
- Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**: 37–46.
- Cohen, J. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* **70**: 213–220.
- Cohen, A., and H. B. Sackrowitz. 1992. An evaluation of some tests of trend in contingency tables. *J.*

- Am. Stat. Assoc.* **87**: 470–475.
- Collett, D. 2003. *Modelling Binary Data*, 2nd ed. London: Chapman & Hall.
- Collins, L. M., and S. T. Lanza. 2009. *Latent Class and Latent Transition Analysis*. Hoboken, NJ: Wiley.
- Colombi, R. 1998. A multivariate logit model with marginal canonical association. *Commun. Stat. Theory Methods* **27**: 2953–2971.
- Colombi, R., and S. Giordano. 2012. Graphical models for multivariate Markov chains. *J. Multiv. Anal.* **107**: 90–103.
- Conaway, M. R. 1989. Analysis of repeated categorical measurements with conditional likelihood methods. *J. Am. Stat. Assoc.* **84**: 53–62.
- Congdon, P. 2005. *Bayesian Models for Categorical Data*. Hoboken, NJ: Wiley.
- Consonni, G., and L. La Rocca. 2008. Intrinsic tests for the equality of two correlated proportions. *J. Am. Stat. Assoc.* **103**: 1260–1269.
- Cook, R. D., and S. Weisberg. 1999. *Applied Regression Including Computing and Graphics*. Hoboken, NJ: Wiley.
- Copas, J. 1973. Randomization models for the matched and unmatched 2×2 tables. *Biometrika* **60**: 467–476.
- Copas, J. 1983. Plotting p against *Appl. Stat.* **32**: 25–31.
- Copas, J. 1988. Binary regression models for contaminated data. *J. R. Stat. Soc. B* **50**: 225–265.
- Copas, J., and S. Eguchi. 2010. Likelihood for statistically equivalent models. *J. R. Stat. Soc. B* **72**: 193–217.
- Corcoran, C., L. Ryan, P. Senchaudhuri, C. Mehta, N. Patel, and G. Molenberghs. 2001. An exact trend test for correlated binary data. *Biometrics* **57**: 941–948.
- Cordeiro, G. M., and P. McCullagh. 1991. Bias correction in generalized linear models. *J. R. Stat. Soc. B* **53**: 629–643.
- Cormack, R. M. 1989. Log-linear models for capture–recapture. *Biometrics* **45**: 395–413.
- Cornfield, J. 1951. A method of estimating comparative rates from clinical data: Applications to cancer of the lung, breast and cervix. *J. Natl. Cancer Inst.* **11**: 1269–1275.
- Cornfield, J. 1956. A statistical problem arising from retrospective studies. In *Proceedings 3rd Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman, **4**: 135–148.
- Cornfield, J. 1962. Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: A discriminant function analysis. *Fed. Proc.* **21**(Suppl. 11): 58–61.
- Coull, B. A., and A. Agresti. 1999. The use of mixed logit models to reflect heterogeneity in capture–recapture studies. *Biometrics* **55**: 294–301.
- Coull, B. A., and A. Agresti. 2000. Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. *Biometrics* **56**: 73–80.
- Cox, C. 1984. An elementary introduction to maximum likelihood estimation for multinomial models: Birch's theorem and the delta method. *Am. Stat.* **38**: 283–287.
- Cox, C. 1995. Location-scale cumulative odds models for ordinal data: A generalized non-linear model approach. *Stat. Med.* **14**: 1191–1203.
- Cox, C. 1996. Nonlinear quasi-likelihood models: Applications to continuous proportions. *Comput. Stat. Data An.* **21**: 449–461.
- Cox, D. R. 1958a. The regression analysis of binary sequences. *J. R. Stat. Soc. B* **20**: 215–242.
- Cox, D. R. 1958b. Two further applications of a model for binary regression. *Biometrika* **45**: 562–565.
- Cox, D. R. 1970. *The Analysis of Binary Data* (2nd ed. 1989, by D. R. Cox and E. J. Snell). London: Chapman & Hall.

- Cox, D. R. 1972. The analysis of multivariate binary data. *Appl. Stat.* **21**: 113–120.
- Cox, D. R. 1983. Some remarks on overdispersion. *Biometrika* **70**: 269–274.
- Cox, D. R., and D. V. Hinkley. 1974. *Theoretical Statistics*. London: Chapman & Hall.
- Cramér, H. 1946. *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.
- Cramer, J. S. 2011. *Logit Models from Economics and Other Fields*. Cambridge, UK: Cambridge University Press.
- Cressie, N., and T. R. C. Read. 1984. Multinomial goodness-of-fit tests. *J. R. Stat. Soc. B* **46**: 440–464.
- Cressie, N., and T. R. C. Read. 1989. Pearson X^2 and the loglikelihood ratio statistic G^2 : A comparative review. *Int. Stat. Rev.* **57**: 19–43.
- Crook, J. F., and I. J. Good. 1980. On the application of symmetric Dirichlet distributions and their mixtures to contingency tables, Part II. *Ann. Stat.* **8**: 1198–1218.
- Crouchley, R. 1995. A random-effects model for ordered categorical data. *J. Am. Stat. Assoc.* **90**: 489–498.
- Crowder, M. J. 1978. Beta-binomial ANOVA for proportions. *Appl. Stat.* **27**: 34–37.
- Crowder, M. 1995. On the use of a working correlation matrix in using generalized linear models for repeated measures. *Biometrika* **82**: 407–410.
- Dahinden, C., M. Kalisch, and P. Bühlmann. 2010. Decomposition and model selection for large contingency tables. *Biomet. J.* **52**: 233–252.
- Daniels, M. J., and C. Gatsonis. 1997. Hierarchical polytomous regression models with applications to health services research. *Stat. Med.* **16**: 2311–2325.
- Daniels, M. J., and C. Gatsonis. 1999. Hierarchical generalized linear models in the analysis of variations in health care utilization. *J. Am. Stat. Assoc.* **94**: 29–42.
- Darroch, J. N. 1962. Interactions in multi-factor contingency tables. *J. R. Stat. Soc. B* **24**: 251–263.
- Darroch, J. N. 1981. The Mantel–Haenszel test and tests of marginal symmetry; fixed-effects and mixed models for a categorical response. *Int. Stat. Rev.* **49**: 285–307.
- Darroch, J. N., and P. I. McCloud. 1986. Category distinguishability and observer agreement. *Austral. J. Stat.* **28**: 371–388.
- Darroch, J. N., and D. Ratcliff. 1972. Generalized iterative scaling for log-linear models. *Ann. Math. Stat.* **43**: 1470–1480.
- Darroch, J. N., S. L. Lauritzen, and T. P. Speed. 1980. Markov fields and log-linear interaction models for contingency tables. *Ann. Stat.* **8**: 522–539.
- Darroch, J. N., S. E. Fienberg, G. F. V. Glonek, and B. W. Junker. 1993. A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *J. Am. Stat. Assoc.* **88**: 1137–1148.
- Das Gupta, S., and M. D. Perlman. 1974. Power of the noncentral F -test: Effect of additional variates on Hotelling's T^2 -test. *J. Am. Stat. Assoc.* **69**: 174–180.
- DasGupta, A., and H. Rubin. 2005. Estimation of binomial parameters when both n, p are unknown. *J. Stat. Plan. Infer.* **130**: 391–404.
- DasGupta, A., and T. Zhang. 2004. Binomial and multinomial parameters, inference on. In *Encyclopedia of Statistical Sciences* Hoboken, NJ: Wiley.
- David, H. A. 1988. *The Method of Paired Comparisons*, 2nd ed. Oxford: Oxford University Press.
- Davis, L. J. 1986a. Exact tests for 2 by 2 contingency tables. *Am. Stat.* **40**: 139–141.
- Davis, L. J. 1986b. Relationship between strictly collapsible and perfect tables. *Statist. Probab. Lett.* **4**: 119–122.
- Davis, L. J. 1989. Intersection union tests for strict collapsibility in three-dimensional contingency tables. *Ann. Stat.* **17**: 1693–1708.

- Davison, A. C. 1991. Residuals and diagnostics. In *Statistical Theory and Modelling. In Honour of Sir David Cox, FRS*, eds. D. V. Hinkley, N. Reid, and E. J. Snell. London: Chapman & Hall, pp. 83–106.
- Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge, UK: Cambridge University Press.
- Davison, A. C., D. A. S. Fraser, and N. Reid. 2006. Improved likelihood inference for discrete data. *J. R. Stat. Soc. B* **68**: 495–508.
- Dawid, A. P., and S. L. Lauritzen. 1993. Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Stat.* **21**: 1272–1317.
- Dawson, R. B., Jr. 1954. A simplified expression for the variance of the χ^2 -function on a contingency table. *Biometrika* **41**: 280.
- Day, N. E., and D. P. Byar. 1979. Testing hypotheses in case-control studies: Equivalence of Mantel-Haenszel statistics and logit score tests. *Biometrics* **35**: 623–630.
- Dellaportas, P., and J. Forster. 1999. Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86**: 615–633.
- Deming, W. E., and F. F. Stephan. 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Stat.* **11**: 427–444.
- de Rooij, M. 2008. The analysis of change, Newton's law of gravity and association models. *J. R. Stat. Soc. A* **171**: 137–157.
- de Rooij, M., and W. Heiser. 2005. Graphical representations and odds ratios in a distance-association model for the analysis of cross-classified data. *Psychometrika* **70**: 99–122.
- Dersimonian, R., and N. Laird. 1986. Meta-analysis in clinical trials. *Controlled Clin. Trials* **7**: 177–188.
- Deville, J.-C., C.-E. Särndal, and O. Sautory. 1993. Generalized raking procedures in survey sampling. *J. Am. Stat. Assoc.* **88**: 1013–1020.
- Dey, D. K., S. K. Ghosh, and B. K. Mallick (eds.). 2000. *Generalized Linear Models: A Bayesian Perspective*. New York: Marcel Dekker.
- Diaconis, P., and B. Efron. 1985. Testing for independence in a two-way table: New interpretations of the chi-square statistic. *Ann. Stat.* **13**: 845–874.
- Diaconis, P., and B. Sturmfels. 1998. Algebraic algorithms for sampling from conditional distributions. *Ann. Stat.* **26**: 363–397.
- Diggle, P. J., P. Heagerty, K.-Y. Liang, and S. L. Zeger. 2002. *Analysis of Longitudinal Data*, 2nd ed. Oxford: Clarendon Press.
- Dittrich, R., R. Hatzinger, and W. Katzenbeisser. 1998. Modeling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Appl. Stat.* **47**: 511–525.
- Dittrich, R., B. Francis, R. Hatzinger, and W. Katzenbeisser. 2007. A paired comparison approach for the analysis of sets of Likert-scale responses. *Stat. Modelling* **7**: 3–28.
- Dobra, A. 2003. Markov bases for decomposable graphical models. *Bernoulli* **9**: 1093–1108.
- Dobra, A., and S. E. Fienberg. 2000. Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proc. Natl. Acad. Sci. USA* **97**: 11885–11892.
- Dobra, A., S. E. Fienberg, A. Rinaldo, A. Slavkovic, and Y. Zhou. 2009. Algebraic statistics and contingency table problems: Log-linear models, likelihood estimation, and disclosure limitation. In *Emerging Applications of Algebraic Geometry*, eds. M. Putinar and S. Sullivant New York: Springer.
- Doksum, K. A., and M. Gasko. 1990. On a correspondence between models in binary regression analysis and in survival analysis. *Int. Stat. Rev.* **58**: 243–252.

- Dong, J. 2005. Simpson's paradox. In *Encyclopedia of Biostatistics*, 2nd ed. Chichester, UK: Wiley, pp. 4947–4948.
- Dong, J., and J. S. Simonoff. 1994. The construction and properties of boundary kernels for smoothing sparse multinomials. *J. Comput. Graph. Stat.* **3**: 57–66.
- Donner, A., and W. W. Hauck. 1986. The large-sample efficiency of the Mantel–Haenszel estimator in the fixed-strata case. *Biometrics* **42**: 537–545.
- Doolittle, M. H. 1888. Association ratios. *Bull. Philos. Soc. Washington* **10**: 83–87, 94–96.
- Drost, F. C., W. C. M. Kallenberg, D. S. Moore, and J. Oosterhoff. 1989. Power approximations to multinomial tests of fit. *J. Am. Stat. Assoc.* **84**: 130–141.
- Drton, M., and T. S. Richardson. 2008. Binary models for marginal independence. *J. R. Stat. Soc. B* **70**: 287–309.
- Ducharme, G. R., and Y. Lepage. 1986. Testing collapsibility in contingency tables. *J. R. Stat. Soc. B* **48**: 197–205.
- Dudbridge, F. 2007. Family-based association. In *Handbook of Statistical Genetics*, 3rd ed., vol. 2, eds. D. J. Balding, M. Bishop, and C. Cannings. Hoboken, NJ: Wiley, pp. 1264–1285.
- Dudoit, S., J. Fridlyand, and T. P. Speed. 2002. Comparison of discriminant methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* **97**: 77–87.
- Dudoit, S., J. P. Shaffer, and J. C. Boldrick. 2003. Multiple hypothesis testing in microarray experiments. *Stat. Sci.* **18**: 71–103.
- Dupont, W. D. 1986. Sensitivity of Fisher's exact test to minor perturbations in 2×2 contingency tables. *Stat. Med.* **5**: 629–635.
- Dyke, G. V., and H. D. Patterson. 1952. Analysis of factorial arrangements when the data are proportions. *Biometrics* **8**: 1–12.
- Edwardes, M. D. deB. 1997. Univariate random cut-points theory for the analysis of ordered categorical data. *J. Am. Stat. Assoc.* **92**: 1114–1123.
- Edwards, A. W. F. 1963. The measure of association in a 2×2 table. *J. R. Stat. Soc. A* **126**: 109–114.
- Edwards, D. 2000. *Introduction to Graphical Modelling*, 2nd ed. New York: Springer-Verlag.
- Edwards, D., and S. Kreiner. 1983. The analysis of contingency tables by graphical models. *Biometrika* **70**: 553–565.
- Efron, B. 1975. The efficiency of logistic regression compared to normal discriminant analysis. *J. Am. Stat. Assoc.* **70**: 892–898.
- Efron, B. 1978. Regression and ANOVA with zero–one data: Measures of residual variation. *J. Am. Stat. Assoc.* **73**: 113–121.
- Efron, B. 1996. Empirical Bayes methods for combining likelihoods. *J. Am. Stat. Assoc.* **91**: 538–550.
- Efron, B., and D. V. Hinkley. 1978. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65**: 457–482.
- Efron, B., and C. Morris. 1975. Data analysis using Stein's estimator and its generalizations. *J. Am. Stat. Assoc.* **70**: 311–319.
- Eguchi, S., and J. Copas. 2002. A class of logistic-type discriminant functions. *Biometrika* **89**: 1–22.
- Ekholm, A., J. W. McDonald, and P. W. F. Smith. 2000. Association models for a multivariate binary response. *Biometrics* **56**: 712–718.
- Emerson, J. D., D. C. Hoaglin, and F. Mosteller. 1993. A comparison of procedures for combining risk differences in sets of 2×2 tables from clinical trials. *J. Ital. Stat. Soc.* **2**: 269–290.
- Eriksson, N., S. E. Fienberg, A. Rinaldo, and S. Sullivant. 2006. Polyhedral conditions for the nonexistence of the mle for hierarchical log-linear models. *J. Symbolic Comput.* **41**: 222–233.

- Erosheva, E., S. Fienberg, and C. Joutard. 2007. Describing disability through individual-level mixture models for multivariate binary data. *Ann. Appl. Stat.* **1**: 502–537.
- Escoufier, Y. 1982. L'analyse des tableaux de contingence simples et multiples. In *Proceedings International Meeting on the Analysis of Multidimensional Contingency Tables* (Rome, 1981), ed. R. Coppi. *Metron* **40**: 53–77.
- Espeland, M. A., and S. L. Handelman. 1989. Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics* **45**: 587–599.
- Espeland, M. A., and S. L. Hui. 1987. A general approach to analyzing epidemiologic data that contain misclassification errors. *Biometrics* **43**: 1001–1012.
- Everitt, B. S., S. Landau, M. Leese, and D. Stahl. 2011. *Cluster Analysis*, 5th ed. Hoboken, NJ: Wiley.
- Fagerland, M. W., S. Lydersen, and P. Laake. 2012. Recommended confidence intervals for two independent binomial proportions. *Stat. Methods Med. Res.* **21**: to appear.
- Fahrmeir, L., and H. Kaufmann. 1987. Regression models for non-stationary categorical time series. *J. Time Ser. Anal.* **8**: 147–160.
- Fahrmeir, L., and G. Tutz. 2001. *Multivariate Statistical Modelling based on Generalized Linear Models*, 2nd ed. New York: Springer-Verlag.
- Fan, J., and Y. Fan. 2008. High dimensional classification using features annealed independence rules. *Ann. Stat.* **36**: 2605–2637.
- Fan, J., and J. Lv. 2010. A selective overview of variable selection in high dimensional feature space. *Stat. Sin.* **20**: 101–148.
- Fan, X., Y. Yuan, and J. S. Liu. 2010. The EM algorithm and the rise of computational biology. *Stat. Sci.* **25**: 476–491.
- Farcomeni, A. 2008. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Stat. Methods Med. Res.* **17**: 347–388.
- Farewell, V. T. 1979. Some results on the estimation of logistic models based on retrospective data. *Biometrika* **66**: 27–32.
- Farrington, C. P., and G. Manning. 1990. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Stat. Med.* **9**: 1447–1454.
- Fay, M. P. 2010a. Two-sided exact tests and matching confidence intervals for discrete data. *R Journal* **2**: 53–58.
- Fay, M. P. 2010b. Confidence intervals that match Fisher's exact or Blaker's exact tests. *Biostatistics* **11**: 373–374.
- Fay, R. 1985. A jackknifed chi-squared test for complex samples. *J. Am. Stat. Assoc.* **80**: 148–157.
- Fienberg, S. E. 1970a. An iterative procedure for estimation in contingency tables. *Ann. Math. Stat.* **41**: 907–917.
- Fienberg, S. E. 1970b. Quasi-independence and maximum likelihood estimation in incomplete contingency tables. *J. Am. Stat. Soc.* **65**: 1610–1616.
- Fienberg, S. E. 1972. The analysis of incomplete multi-way contingency tables. *Biometrics* **28**: 177–202.
- Fienberg, S. E. 1980. Fisher's contributions to the analysis of categorical data. In *R. A. Fisher: An Appreciation*, eds. S. E. Fienberg and D. V. Hinkley. Berlin: Springer-Verlag, pp. 75–84.
- Fienberg, S. E. 1984. The contributions of William Cochran to categorical data analysis. In *W. G. Cochran's Impact on Statistics*, eds. P. S. R. S. Rao and J. Sedransk. New York: Wiley, pp. 103–118.
- Fienberg, S. E. 2011. Chi-squared tests and log-linear models to models of mixed membership. *Stat. Biopharm. Res.* **3**: 173–184.

- Fienberg, S. E., and P. W. Holland. 1973. Simultaneous estimation of multinomial cell probabilities. *J. Am. Stat. Assoc.* **68**: 683–690.
- Fienberg, S. E., and K. Larntz. 1976. Loglinear representation for paired and multiple comparison models. *Biometrika* **63**: 245–254.
- Fienberg, S. E., and U. E. Makov. 1998. Confidentiality, uniqueness, and disclosure limitation for categorical data. *J. Official Stat.* **14**: 385–397.
- Fienberg, S. E., and A. Rinaldo. 2007. Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *J. Stat. Plan. Infer.* **137**: 3430–3445.
- Fienberg, S. E., and A. Rinaldo. 2011. Maximum likelihood estimation in log-linear models. arXiv:1104.3618v1 [math.ST]. Submitted for publication.
- Fienberg, S. E., M. A. Johnson, and B. J. Junker, 1999. Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *J. R. Stat. Soc. A* **162**: 383–405.
- Finney, D. J. 1947. The estimation from individual records of the relationship between dose and quantal response. *Biometrika* **34**: 320–334.
- Finney, D. J. 1971. *Probit Analysis*, 3rd ed. Cambridge, UK: Cambridge University Press.
- Firth, D. 1987. On the efficiency of quasi-likelihood estimation. *Biometrika* **74**: 233–245.
- Firth, D. 1993a. Bias reduction of maximum likelihood estimates. *Biometrika* **80**: 27–38.
- Firth, D. 1993b. Recent developments in quasi-likelihood methods. *Proc. ISI 49th Session*, pp. 341–358.
- Firth, D., and R. X. De Menezes. 2004. Quasi-variances. *Biometrika* **91**: 65–80.
- Fischer, G. H., and I. W. Molenaar. 1995. *Rasch Models: Foundations, Recent Developments, and Applications*. New York: Springer-Verlag.
- Fisher, R. A. 1922. On the interpretation of chi-square from contingency tables, and the calculation of P . *J. R. Stat. Soc.* **85**: 87–94.
- Fisher, R. A. 1924. The conditions under which chi-square measures the discrepancy between observation and hypothesis. *J. R. Stat. Soc.* **87**: 442–450.
- Fisher, R. A. 1926. Bayes' theorem and the fourfold table. *Eugenics Rev.* **18**: 32–33.
- Fisher, R. A. 1934, 1970. *Statistical Methods for Research Workers* (originally published 1925, 14th ed., 1970.) Edinburgh: Oliver and Boyd.
- Fisher, R. A. 1935a. *The Design of Experiments* (8th ed., 1966). Edinburgh: Oliver & Boyd.
- Fisher, R. A. 1935b. Appendix to article by C. Bliss. *Ann. Appl. Biol.* **22**: 164–165.
- Fisher, R. A. 1935c. The logic of inductive inference. *J. R. Stat. Soc.* **98**: 39–82.
- Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7**: 179–188.
- Fisher, R. A. 1945. A new test for 2×2 tables (Letter to the Editor). *Nature* **156**: 388.
- Fisher, R. A. 1954. The analysis of variance with various binomial transformations. *Biometrics* **10**: 130–139.
- Fisher, R. A. 1956. *Statistical Methods for Scientific Inference*. Edinburgh: Oliver & Boyd.
- Fisher, R. A., and F. Yates. 1938. *Statistical Tables*. Edinburgh: Oliver & Boyd.
- Fitzmaurice, G. M., and N. M. Laird. 1993. A likelihood-based method for analysing longitudinal binary responses. *Biometrika* **80**: 141–151.
- Fitzmaurice, G. M., N. M. Laird, and A. G. Rotnitzky. 1993. Regression models for discrete longitudinal responses. *Stat. Sci.* **8**: 284–299.
- Fitzmaurice, G. M., N. M. Laird, and S. Lipsitz. 1994. Analysing incomplete longitudinal binary responses: A likelihood-based approach. *Biometrics* **50**: 601–612.
- Fitzpatrick, S., and A. Scott. 1987. Quick simultaneous confidence intervals for multinomial

- proportions. *J. Am. Stat. Assoc.* **82**: 875–878.
- Fleiss, J. L. 1982. A simplification of the classic large-sample standard error of a function of multinomial proportions. *Am. Stat.* **36**: 377–378.
- Fleiss, J. L., and J. Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Meas.* **33**: 613–619.
- Fleiss, J. L., J. Cohen, and B. S. Everitt. 1969. Large-sample standard errors of kappa and weighted kappa. *Psychol. Bull.* **72**: 323–327.
- Fleiss, J. L., B. Levin, and M. C. Paik. 2003. *Statistical Methods for Rates and Proportions*. 3rd ed. Hoboken, NJ: Wiley.
- Fokianos, K., and B. Kedem. 2002. *Regression Models for Time Series Data*. Hoboken, NJ: Wiley.
- Fokianos, K., and B. Kedem. 2003. Regression theory for categorical time series. *Stat. Sci.* **18**: 357–376.
- Follman, D. A., and D. Lambert. 1989. Generalizing logistic regression by nonparametric mixing. *J. Am. Stat. Assoc.* **84**: 295–300.
- Forster, J. J. 2010. Bayesian inference for Poisson and multinomial log-linear models. *Stat. Methodol.* **7**: 210–224.
- Forster, J. J., and P. W. F. Smith. 1998. Model-based inference for categorical survey data subject to non-ignorable non-response. *J. R. Stat. Soc. B* **60**: 57–70.
- Forster, J. J., J. W. McDonald, and P. W. F. Smith. 1996. Monte Carlo exact conditional tests for log-linear and logistic models. *J. R. Stat. Soc. B* **58**: 445–453.
- Fowlkes, E. B. 1987. Some diagnostics for binary logistic regression via smoothing. *Biometrika* **74**: 503–515.
- Fowlkes, E. B., A. E. Freeny, and J. Landwehr. 1988. Evaluating logistic models for large contingency tables. *J. Am. Stat. Assoc.* **83**: 611–622.
- Fraley, C., and A. E. Raftery. 2002. Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.* **97**: 611–631.
- Freeman, G. H., and J. H. Halton. 1951. Note on an exact treatment of contingency, goodness-of-fit and other problems of significance. *Biometrika* **38**: 141–149.
- Freeman, M. F., and J. W. Tukey. 1950. Transformations related to the angular and the square root. *Ann. Math. Stat.* **21**: 607–611.
- Freidlin, B., and J. L. Gastwirth. 1999. Unconditional versions of several tests commonly used in the analysis of contingency tables. *Biometrics* **55**: 264–267.
- Friedman, J. H. 1989. Regularized discriminant analysis. *J. Am. Stat. Assoc.* **84**: 165–175.
- Friedman, J. H., and J. J. Meulman. 2004. Clustering objects on subsets of attributes. *J. R. Stat. Soc. B* **66**: 815–849.
- Friendly, M. 1994. Mosaic displays for multi-way contingency tables. *J. Am. Stat. Assoc.* **89**: 190–200.
- Frome, E. L. 1983. The analysis of rates using Poisson regression models. *Biometrics* **39**: 665–674.
- Gabriel, K. R. 1966. Simultaneous test procedures for multiple comparisons on categorical data. *J. Am. Stat. Assoc.* **61**: 1081–1096.
- Gabriel, K. R. 1971. The biplot graphic display of matrices with applications to principal component analysis. *Biometrika* **58**: 453–467.
- Gail, M., and N. Mantel. 1977. Counting the number of $r \times c$ contingency tables with fixed margins. *J. Am. Stat. Assoc.* **72**: 859–862.
- Gart, J. J. 1969. An exact test for comparing matched proportions in crossover designs. *Biometrika* **56**: 75–80.
- Gart, J. J. 1970. Point and interval estimation of the common odds ratio in the combination of 2×2

- tables with fixed margins. *Biometrika* **57**: 471–475.
- Gart, J. J. 1971. The comparison of proportions: A review of significance tests, confidence intervals and adjustments for stratification. *Rev. Int. Stat. Rev.* **39**: 148–169.
- Gart, J. J., and J. Nam. 1988. Approximate interval estimation of the ratio of binomial parameters: A review and corrections for skewness. *Biometrics* **44**: 323–338.
- Gart, J. J., and J. R. Zweifel. 1967. On the bias of various estimators of the logit and its variance with applications to quantal bioassay. *Biometrika* **54**: 181–187.
- Gelfand, A. E., and A. F. M. Smith. 1990. Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* **85**: 398–409.
- Gelman, A., and J. Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. *Bayesian Data Analysis*, 2nd ed. Boca Raton, FL: CRC Press.
- Geng, Z. 1992. Collapsibility of relative risk in contingency tables with a response variable. *J. R. Stat. Soc. B* **54**: 585–593.
- Genkin, A., D. D. Lewis, and D. Madigan. 2007. Large-scale Bayesian logistic regression for text categorization. *Technometrics* **49**: 291–304.
- Genter, F. C., and V. T. Farewell. 1985. Goodness-of-link testing in ordinal regression models. *Can. J. Stat.* **13**: 37–44.
- Geyer, C. J., and G. D. Meeden. 2005. Fuzzy and randomized confidence intervals and *P*-values. *Stat. Sci.* **20**: 358–366.
- Ghosh, B. K. 1979. A comparison of some approximate confidence intervals for the binomial parameter. *J. Am. Stat. Assoc.* **74**: 894–900.
- Ghosh, M., and B. Mukherjee. 2010. Bayesian analysis of matched pair data. In *Frontiers of Statistical Decision Making and Bayesian Analysis*, eds. M.-H. Chen, D. K. Dey, P. Müller, D. Sun, and K. Ye. New York: Springer, pp. 436–445.
- Ghosh, M., M.-H. Chen, A. Ghosh, and A. Agresti. 2000. Hierarchical Bayesian analysis of binary matched pairs data. *Stat. Sin.* **10**: 647–657.
- Gibbons, R. D., and D. Hedeker. 1997. Random-effects probit and logistic regression models for three-level data. *Biometrics* **53**: 1527–1537.
- Gilbert, P. B. 2005. A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Appl. Stat.* **54**: 143–158.
- Gilmour, A. R., R. D. Anderson, and A. L. Rae. 1985. The analysis of binomial data by a generalized linear mixed model. *Biometrika* **72**: 593–599.
- Gilula, Z. 1984. On some similarities between canonical correlation models and latent class models for two-way contingency tables. *Biometrika* **71**: 523–529.
- Gilula, Z., and S. Haberman. 1986. Canonical analysis of contingency tables by maximum likelihood. *J. Am. Stat. Assoc.* **83**: 780–788.
- Gilula, Z., and S. Haberman. 1988. The analysis of multivariate contingency tables by restricted canonical and restricted association models. *J. Am. Stat. Assoc.* **83**: 760–771.
- Gilula, Z., and S. Haberman. 2005. Chi-square, partition of. In *Encyclopedia of Biostatistics*, 2nd ed. Chichester, UK: Wiley, pp. 760–765.
- Gini, C. 1914a. Sulla misura della concentrazione e della variabilità dei caratteri, *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti*. **73**: 1203–1248. (English translation in *Metron* **63**: 3–38, 2005.)
- Gini, C. 1914b. Di una misura della dissomiglianza tra due gruppi di quantità e delle sue applicazioni allo studio delle relazioni statistiche. *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti*. **74**: 185–213.

- Gleser, L. J., and D. S. Moore. 1985. The effect of positive dependence on chi-squared tests for categorical data. *J. R. Stat. Soc. B* **47**: 459–465.
- Glonek, G. 1996. A class of regression models for multivariate categorical responses. *Biometrika* **83**: 15–28.
- Glonek, G. F. V., and P. McCullagh. 1995. Multivariate logistic models. *J. R. Stat. Soc. B* **57**: 533–546.
- Glonek, G., J. N. Darroch, and T. P. Speed. 1988. On the existence of maximum likelihood estimators for hierarchical loglinear models. *Scand. J. Stat.* **15**: 187–193.
- Gokhale, D. V., and S. Kullback. 1978. *The Information in Contingency Tables*. New York: Marcel Dekker.
- Goldstein, H. 2010. *Multilevel Statistical Models*, 4th ed. Hoboken, NJ: Wiley.
- Good, I. J. 1956. On the estimation of small frequencies in contingency tables. *J. R. Stat. Soc. B* **18**: 113–124.
- Good, I. J. 1963. Maximum entropy for hypothesis formulation, especially for multi-dimensional contingency tables. *Ann. Math. Stat.* **34**: 911–934.
- Good, I. J. 1965. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge, MA: MIT Press.
- Good, I. J. 1967. A Bayesian significance test for multinomial distributions. *J. R. Stat. Soc. B* **29**: 399–431.
- Good, I. J. 1976. On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Ann. Stat.* **4**: 1159–1189.
- Good, I. J., and R. A. Gaskins. 1971. Nonparametric roughness penalties for probability densities. *Biometrika* **58**: 255–277.
- Good, I. J., and Y. Mittal. 1987. The amalgamation and geometry of two-by-two contingency tables. *Ann. Stat.* **15**: 694–711.
- Good, I. J., T. N. Gover, and G. J. Mitchell. 1970. Exact distributions for χ^2 and for the likelihood-ratio statistic for the equiprobable multinomial distribution. *J. Am. Stat. Assoc.* **65**: 267–283.
- Goodman, L. A. 1964a. Simultaneous confidence intervals for cross-product ratios in contingency tables. *J. R. Stat. Soc. B* **26**: 86–102.
- Goodman, L. A. 1964b. Interactions in multi-dimensional contingency tables. *Ann. Math. Stat.* **35**: 632–646.
- Goodman, L. A. 1965. On simultaneous confidence intervals for multinomial proportions. *Technometrics* **7**: 247–254.
- Goodman, L. A. 1968. The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries. *J. Am. Stat. Assoc.* **63**: 1091–1131.
- Goodman, L. A. 1969. On partitioning chi-square and detecting partial association in three-way contingency tables. *J. R. Stat. Soc. B* **31**: 486–498.
- Goodman, L. A. 1970. The multivariate analysis of qualitative data: Interaction among multiple classifications. *J. Am. Stat. Assoc.* **65**: 226–256.
- Goodman, L. A. 1971a. The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics* **13**: 33–61.
- Goodman, L. A. 1971b. The partitioning of chi-square, the analysis of marginal contingency tables, and the estimation of expected frequencies in multidimensional contingency tables. *J. Am. Stat. Assoc.* **66**: 339–344.
- Goodman, L. A. 1973. The analysis of multidimensional contingency tables when some variables are posterior to others: A modified path analysis approach. *Biometrika* **60**: 179–192.

- Goodman, L. A. 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61**: 215–231.
- Goodman, L. A. 1979a. Simple models for the analysis of association in cross-classifications having ordered categories. *J. Am. Stat. Assoc.* **74**: 537–552.
- Goodman, L. A. 1979b. Multiplicative models for square contingency tables with ordered categories. *Biometrika* **66**: 413–418.
- Goodman, L. A. 1981a. Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *J. Am. Stat. Assoc.* **76**: 320–334.
- Goodman, L. A. 1981b. Association models and the bivariate normal for contingency tables with ordered categories. *Biometrika* **68**: 347–355.
- Goodman, L. A. 1983. The analysis of dependence in cross-classifications having ordered categories, using log-linear models for frequencies and log-linear models for odds. *Biometrics* **39**: 149–160.
- Goodman, L. A. 1985. The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Ann. Stat.* **13**: 10–69.
- Goodman, L. A. 1986. Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *Int. Stat. Rev.* **54**: 243–309.
- Goodman, L. A. 1996. A single general method for the analysis of cross-classified data: Reconciliation and synthesis of some methods of Pearson, Yule, and Fisher, and also some methods of correspondence analysis and association analysis. *J. Am. Stat. Assoc.* **91**: 408–427.
- Goodman, L. A. 2000. The analysis of cross-classified data: Notes on a century of progress in contingency table analysis, and some comments on its prehistory and its future. In *Statistics for the 21st Century*, eds. C. R. Rao and G. J. Székely. New York: Marcel Dekker, pp. 189–231.
- Goodman, L. A. 2007. Statistical magic and/or statistical serendipity: An age of progress in the analysis of statistical data. *Annu. Rev. Sociol.* **33**: 1–19.
- Goodman, L. A., and W. H. Kruskal. 1979. *Measures of Association for Cross Classifications*. New York: Springer-Verlag (contains articles appearing in *J. Am. Stat. Assoc.* in 1954, 1959, 1963, 1972).
- Gottard, A., G. M. Marchetti, and A. Agresti. 2011. Quasi-symmetric graphical log-linear models. *Scand. J. Stat.* **38**: 447–465.
- Gourieroux, C., A. Monfort, and A. Trognon. 1984. Pseudo maximum likelihood methods: Theory. *Econometrica* **52**: 681–700.
- Graubard, B. I., and E. L. Korn. 1987. Choice of column scores for testing independence in ordered $2 \times K$ contingency tables. *Biometrics* **43**: 471–476.
- Green, P. J. 1984. Iteratively weighted least squares for maximum likelihood estimation and some robust and resistant alternatives. *J. R. Stat. Soc. B* **46**: 149–192.
- Greenacre, M. A. 1988. Correspondence analysis of multivariate categorical data by weighted least squares. *Biometrika* **75**: 457–467.
- Greenacre, M. J. 2007. *Correspondence Analysis in Practice*, 2nd ed. Boca Raton, FL: Chapman & Hall/CRC Press.
- Greene, W. H., and D. A. Hensher. 2010. *Modeling Ordered Choices: A Primer*. Cambridge, UK: Cambridge University Press.
- Greenhouse, J. B. 2009. Commentary: Cornfield, epidemiology and causality. *Int. J. Epidemiol.* **38**: 1199–1201.
- Greenland, S. 1991. On the logical justification of conditional tests for two-by-two contingency tables. *Am. Stat.* **45**: 248–251.
- Greenland, S., and J. M. Robins. 1985. Estimation of a common effect parameter from sparse follow-

- up data. *Biometrics* **41**: 55–68.
- Greenland, S., J. M. Robins, and J. Pearl. 1999. Confounding and collapsibility in causal inference. *Stat. Sci.* **14**: 29–46.
- Greenwood, M., and G. U. Yule. 1920. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *J. R. Stat. Soc. A* **83**: 255–279.
- Greenwood, P. E., and M. S. Nikulin. 1996. *A Guide to Chi-Squared Testing*. Hoboken, NJ: Wiley.
- Grevstad, N. 2006. Binomial distribution: Sample size estimation. In *Encyclopedia of Statistical Sciences*. Hoboken, NJ: Wiley.
- Grizzle, J. E., C. F. Starmer, and G. G. Koch. 1969. Analysis of categorical data by linear models. *Biometrics* **25**: 489–504.
- Gross, S. T. 1981. On asymptotic power and efficiency of tests of independence in contingency tables with ordered classifications. *J. Am. Stat. Assoc.* **76**: 935–941.
- Gueorguieva, R. 2001. A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Stat. Modelling* **1**: 177–193.
- Gueorguieva, R., and A. Agresti. 2001. A correlated probit model for joint modeling of clustered binary and continuous responses. *J. Am. Stat. Assoc.* **96**: 1102–1112.
- Guerrero, V. M., and R. A. Johnson. 1982. Use of the Box–Cox transformation with binary response models. *Biometrika* **69**: 309–314.
- Guirnarães, P. 2005. A simple approach to estimate the beta-binomial model. *Stata J.* **5**: 385–394.
- Guirnarães, P., and R. C. Lindrooth. 2007. Controlling for overdispersion in grouped conditional logit models: A computationally simple application of Dirichlet-multinomial regression. *Econometrics J.* **10**: 439–452.
- Gunel, E., and J. Dickey. 1974. Bayes factors for independence in contingency tables. *Biometrika* **61**: 545–557.
- Guo, G., and Zhao, H. 2000. Multilevel modeling for binary data. *Annu. Rev. Sociol.* **26**: 441–462.
- Gurland, J., I. Lee, and P. A. Dahm. 1960. Polychotomous quantal response in biological assay. *Biometrics* **16**: 382–398.
- Haber, M. 1985. Maximum likelihood methods for linear and log-linear models in categorical data. *Comput. Stat. Data An.* **3**: 1–10.
- Haber, M. 1986. An exact unconditional test for the 2×2 comparative trial. *Psychol. Bull.* **99**: 129–132.
- Haber, M. 1989. Do the marginal totals of a 2×2 contingency table contain information regarding the table proportions? *Commun. Stat. A* **18**: 147–156.
- Haberman, S. J. 1973a. The analysis of residuals in cross-classification tables. *Biometrics* **29**: 205–220.
- Haberman, S. J. 1973b. Log-linear models for frequency data: Sufficient statistics and likelihood equations. *Ann. Stat.* **1**: 617–632.
- Haberman, S. J. 1974a. *The Analysis of Frequency Data*. Chicago: University of Chicago Press.
- Haberman, S. J. 1974b. Log-linear models for frequency tables with ordered classifications. *Biometrics* **36**: 589–600.
- Haberman, S. J. 1977a. Log-linear models and frequency tables with small expected cell counts. *Ann. Stat.* **5**: 1148–1169.
- Haberman, S. J. 1977b. Maximum likelihood estimation in exponential response models. *Ann. Stat.* **5**: 815–841.
- Haberman, S. J. 1978, 1979. *Analysis of Qualitative Data*, Vols. 1 and 2. New York: Academic Press.

- Haberman, S. J. 1981. Tests for independence in two-way contingency tables based on canonical correlation and on linear-by-linear interaction. *Ann. Stat.* **9**: 1178–1186.
- Haberman, S. J. 1982. The analysis of dispersion of multinomial responses. *J. Am. Stat. Assoc.* **77**: 568–580.
- Haberman, S. J. 1988. A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *J. Am. Stat. Assoc.* **83**: 555–560.
- Haberman, S. J. 1995. Computation of maximum likelihood estimates in association models. *J. Am. Stat. Assoc.* **90**: 1438–1446.
- Hagenaars, J. A., and A. L. McCutcheon (eds.). 2009. *Applied Latent Class Analysis*. Cambridge, UK: Cambridge University Press.
- Hald, A. 1998. *A History of Mathematical Statistics from 1750 to 1930*. Hoboken, NJ: Wiley.
- Haldane, J. B. S. 1940. The mean and variance of χ^2 , when used as a test of homogeneity, when expectations are small. *Biometrika* **31**: 346–355.
- Haldane, J. B. S. 1948. The precision of observed values of small frequencies. *Biometrika* **35**: 297–303.
- Haldane, J. B. S. 1956. The estimation and significance of the logarithm of a ratio of frequencies. *Ann. Human Genet.* **20**: 309–311.
- Hall, P., and D. M. Titterington. 1987. On smoothing sparse multinomial data. *Austral. J. Stat.* **29**: 19–37.
- Hamada, M., and C. F. J. Wu. 1990. A critical look at accumulation analysis and related methods. *Technometrics* **32**: 119–130.
- Hand, D. J. 2006. Classifier technology and the illusion of progress. *Stat. Sci.* **21**: 1–14.
- Hand, D. J. 2009. Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Mach. Learn.* **77**: 103–123.
- Hand, D. J., and W. E. Henley. 1997. Statistical classification methods in consumer credit scoring: A review. *J. R. Stat. Soc. A* **160**: 523–541.
- Hanley, J. A., and B. J. McNeil. 1982. The meaning and use of the area under a receiving operating characteristic (ROC) curve. *Radiology* **143**: 29–36.
- Hansen, L. P. 1982. Large sample properties of generalized-method of moments estimators. *Econometrica* **50**: 1029–1054.
- Hartzel, J., I.-M. Liu, and A. Agresti. 2001a. Describing heterogeneous effects in stratified ordinal contingency tables, with application to multi-center clinical trials. *Comput. Stat. Data An.* **35**: 429–449.
- Hartzel, J., A. Agresti, and B. Caffo. 2001b. Multinomial logit random effects models. *Statist. Modelling* **1**: 81–102.
- Hashemi, L., B. Nandrum, and R. Goldberg. 1997. Bayesian analysis for a single 2×2 table. *Stat. Med.* **16**: 1311–1328.
- Haslett, S. 1990. Degrees of freedom and parameter estimability in hierarchical models for sparse complete contingency tables. *Comput. Stat. Data An.* **9**: 179–195.
- Hastie, T., and R. Tibshirani. 1987. Non-parametric logistic and proportional odds regression. *Appl. Stat.* **36**: 260–276.
- Hastie, T., and R. Tibshirani. 1990. *Generalized Additive Models*. London: Chapman & Hall.
- Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer.
- Hauck, W. W. 1983. A note on confidence bands for the logistic response curve. *Am. Stat.* **37**: 158–160.
- Hauck, W. W., and A. Donner. 1977. Wald's test as applied to hypotheses in logit analysis. *J. Am.*

- Stat. Assoc.* **72**: 851–853.
- Hausman, J. A., and D. A. Wise. 1978. A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica* **46**: 403–426.
- Heagerty, P. J. 1999. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics* **55**: 688–698.
- Heagerty, P. J. 2002. Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics* **58**: 342–351.
- Heagerty, P. J., and S. R. Lele. 1998. A composite likelihood approach to binary spatial data. *J. Am. Stat. Assoc.* **93**: 1099–1111.
- Heagerty, P. J., and S. L. Zeger. 1996. Marginal regression models for clustered ordinal measurements. *J. Am. Stat. Assoc.* **91**: 1024–1036.
- Heagerty, P. J., and S. L. Zeger. 2000. Marginalized multilevel models and likelihood inference. *Stat. Sci.* **15**: 1–19.
- Heckman, J., and B. Singer. 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* **52**: 271–320.
- Hedeker, D. 2008. Multilevel models for ordinal and nominal variables. In *Handbook of Multilevel Analysis*, eds. J. de Lecuw and E. Meijer. New York: Springer, Chap. 6, pp. 239–276.
- Hedeker, D., and R. D. Gibbons. 1994. A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50**: 933–944.
- Hedeker, D., and R. D. Gibbons. 2006. *Longitudinal Data Analysis*. Hoboken, NJ: Wiley.
- Heinen, T. 1996. *Latent Class and Discrete Latent Trait Models*. Thousand Oaks, CA: Sage Publications.
- Heinze, G., and M. Schemper. 2002. A solution to the problem of separation in logistic regression. *Stat. Med.* **21**: 2409–2419.
- Heyde, C. C. 1997. *Quasi-likelihood and Its Application*. New York: Springer-Verlag.
- Hilbe, J. M. 2011. *Negative Binomial Regression*, 2nd ed. Cambridge, UK: Cambridge University Press.
- Hinde, J. 1982. Compound Poisson regression models. In *GLIM82: Proceedings International Conference on Generalised Linear Models*, ed. R. Gilchrist. New York: Springer-Verlag, pp. 109–121.
- Hinde, J., and C. G. B. Demétrio. 1998. Overdispersion: Models and estimation. *Comput. Stat. Data An.* **27**: 151–170.
- Hirji, K. F. 2005. *Exact Analysis of Discrete Data*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Hirji, K. F., C. R. Mehta, and N. R. Patel. 1987. Computing distributions for exact logistic regression. *J. Am. Stat. Assoc.* **82**: 1110–1117.
- Hirotsu, C. 1982. Use of cumulative efficient scores for testing ordered alternatives in discrete models. *Biometrika* **69**: 567–577.
- Hirschfeld, H. O. 1935. A connection between correlation and contingency. *Cambridge Philos. Soc. Proc. (Math. Proc.)* **31**: 520–524.
- Hitchcock, D. B. 2009. Yates and contingency tables: 75 years later. *Electronic J. History Prob. Stat.* **5**: no. 2, 1–14.
- Hitchcock, D. B., and Z. Chen. 2008. Smoothing dissimilarities to cluster binary data. *Comput. Stat. Data An.* **52**: 4699–4711.
- Hodges, J. L., Jr. 1958. Fitting the logistic by maximum likelihood. *Biometrics* **14**: 453–461.
- Hoem, J. M. 1987. Statistical analysis of a multiplicative model and its application to the standardization of vital rates: A review. *Int. Stat. Rev.* **5**: 119–152.
- Hoeting, J. A., M. Leecaster, and D. Bowden. 2000. An improved model for spatially correlated

- binary responses. *J. Agric. Biol. Environ. Stat.* **5**: 102–114.
- Hoff, P. D. 2009. *A First Course in Bayesian Statistical Methods*. New York: Springer.
- Holford, T. R. 1980. The analysis of rates and of survivorship using log-linear models. *Biometrics* **36**: 299–305.
- Holland, P. W., and H. Wainer (eds.) 1993. *Differential Item Functioning*. New York: Routledge.
- Holmes, C. C., and L. Held. 2006. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* **1**: 145–168.
- Holt, D., A. J. Scott, and P. D. Ewings. 1980. Chi-squared tests with survey data. *J. R. Stat. Soc. A* **143**: 303–320.
- Hook, E. B., and R. R. Regal. 1995. Capture–recapture methods in epidemiology: Methods and limitations. *Epidemiol. Rev.* **17**: 243–264.
- Hosmer, D. W., and S. Lemeshow. 1980. A goodness-of-fit test for multiple logistic regression model. *Commun. Stat. A* **9**: 1043–1069.
- Hosmer, D. W., and S. Lemeshow. 2000. *Applied Logistic Regression*, 2nd ed. Hoboken, NJ: Wiley.
- Hosmer, D. W., T. Hosmer, S. le Cessie, and S. Lemeshow. 1997. A comparison of goodness-of-fit tests for the logistic regression model. *Stat. Med.* **16**: 965–980.
- Hout, M., O. D. Duncan, and M. E. Sobel. 1987. Association and heterogeneity: Structural models of similarities and differences. *Sociol. Methodol.* **17**: 145–184.
- Howard, J. V. 1998. The 2×2 table: A discussion from a Bayesian viewpoint. *Stat. Sci.* **13**: 351–367.
- Hsieh, F. Y. 1989. Sample size tables for logistic regression. *Stat. Med.* **8**: 795–802.
- Hsieh, F. Y., D. A. Bloch, and M. D. Larsen. 1998. A simple method of sample size calculation for linear and logistic regression. *Stat. Med.* **17**: 1623–1634.
- Hu, B., M. Palta, and J. Shao. 2005. Properties of R^2 statistics for logistic regression. *Stat. Med.* **25**: 1383–1395.
- Hunt, L. A., and M. A. Jorgensen. 1999. Mixture model clustering using the multimix program. *Austral. New Zealand J. Stat.* **41**: 153–171.
- Hwang, J. T. G., and M.-C. Yang. 2001. An optimality theory for mid P -values in 2×2 contingency tables. *Stat. Sin.* **11**: 807–826.
- Ibrahim, J., and P. W. Laud. 1991. On Bayesian analysis of generalized linear models using Jeffreys's Prior. *J. Am. Stat. Assoc.* **86**: 981–986.
- Ibrahim, J., M.-H. Chen, S. R. Lipsitz, and A. H. Herring. 2005. Missing-data methods for generalized linear models. *J. Am. Stat. Assoc.* **100**: 332–346.
- Imai, K., and D. A. van Dyk. 2005. A Bayesian analysis of the multinomial probit model using marginal data augmentation. *J. Econometrics* **124**: 311–334.
- Imrey, P. B. 2005. Bradley–Terry model. In *Encyclopedia of Biostatistics*, 2nd ed. Chichester, UK: Wiley, pp. 551–558.
- Imrey, P. B. 2011. Koch's contributions to statistical practice, 1965–1985. *Stat. Biopharm. Res.* **3**: 139–162.
- Imrey, P. B., W. D. Johnson, and G. G. Koch. 1976. An incomplete contingency table approach to paired-comparison experiments. *J. Am. Stat. Assoc.* **71**: 614–623.
- Imrey, P. B., G. G. Koch, and M. E. Stokes. 1981. Categorical data analysis: Some reflections on the log linear model and logistic regression. I: Historical and methodological overview. *Int. Stat. Rev.* **49**: 265–283.
- Imrey, P. B., G. G. Koch, and J. S. Preisser. 1996. The evolution of categorical data modeling: A biometric perspective. In *Advances in Biometry*, eds. P. Armitage and H. A. David. Hoboken, NJ: Wiley.

- Ireland, C. T., and S. Kullback. 1968a. Minimum discrimination information estimation. *Biometrics* **24**: 707–713.
- Ireland, C. T., and S. Kullback. 1968b. Contingency tables with given marginals. *Biometrika* **55**: 179–188.
- Ireland, C. T., H. H. Ku, and S. Kullback. 1969. Symmetry and marginal homogeneity of an $r \times r$ contingency table. *J. Am. Stat. Assoc.* **64**: 1323–1341.
- Irwin, J. O. 1935. Tests of significance for differences between percentages based on small numbers. *Metron* **12**: 83–94.
- Jensen, S. T., X. S. Liu, Q. Zhou, and J. S. Liu. 2004. Computational discovery of gene regulatory binding motifs: A Bayesian perspective. *Stat. Sci.* **19**: 188–204.
- Johnson, B. M. 1971. On the admissible estimators for certain fixed sample binomial problems. *Ann. Math. Stat.* **42**: 1579–1587.
- Johnson, N. L., A. W. Kemp, and S. Kotz. 2005. *Univariate Discrete Distributions*, 3rd ed. Hoboken, NJ: Wiley.
- Johnson, W. 1985. Influence measures for logistic regression: Another point of view. *Biometrika* **72**: 59–65.
- Johnson, V. E., and J. H. Albert. 1999. *Ordinal Data Modeling*. New York: Springer.
- Jones, B., and M. G. Kenward. 1987. Modelling binary data from a three-period cross-over trial. *Stat. Med.* **6**: 555–564.
- Jones, M. P., T. W. O’Gorman, J. H. Lemke, and R. F. Woolson. 1989. A Monte Carlo investigation of homogeneity tests of the odds ratio under various sample size considerations. *Biometrics* **45**: 171–181.
- Jørgensen, B. 1983. Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika* **70**: 19–28.
- Jørgensen, B. 1987. Exponential dispersion models. *J. R. Stat. Soc. B* **49**: 127–162.
- Kalbfleisch, J. D., and J. F. Lawless. 1985. The analysis of panel data under a Markov assumption. *J. Am. Stat. Assoc.* **80**: 863–871.
- Kass, G. V. 1980. An exploratory technique for investigating large quantities of categorical data. *Appl. Stat.* **29**: 119–127.
- Kastner, C., A. Fieger, and C. Heumann. 1997. MAREG and WinMAREG: A tool for marginal regression models. *Comput. Stat. Data An.* **24**: 237–241.
- Kateri, M., and A. Agresti. 2007. A class of ordinal quasi-symmetry models for square contingency tables. *Stat. Prob. Lett.* **77**: 598–603.
- Kateri, M., and A. Agresti. 2010. A generalized regression model for a binary response. *Stat. Prob. Lett.* **80**: 89–95.
- Kateri, M., and T. Papaioannou. 1997. Asymmetry models for contingency tables. *J. Am. Stat. Assoc.* **92**: 1124–1131.
- Kateri, M., A. Nicolaou, and I. Ntzoufras. 2005. Bayesian inference for the RC(m) association model. *J. Comput. Graph. Stat.* **14**: 116–138.
- Kauermann, G., and R. J. Carroll. 2001. A note on the efficiency of sandwich covariance matrix estimation. *J. Am. Stat. Assoc.* **96**: 1387–1397.
- Kauermann, G., and G. Tutz. 2001. Testing generalized linear and semiparametric models against smooth alternatives. *J. R. Stat. Soc. B* **63**: 147–166.
- Kaufman, L., and P. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: Wiley.
- Kawaguchi, A., G. G. Koch, and X. Wang. 2011. Stratified multivariate Mann–Whitney estimators for the comparison of two treatments with randomization based covariance adjustment. *Stat.*

- Biopharm. Res.* **3**: 217–231.
- Kelderman, H. 1984. Loglinear Rasch model tests. *Psychometrika* **49**: 223–245.
- Kempthorne, O. 1979. In dispraise of the exact test: Reactions. *J. Stat. Plan. Infer.* **3**: 199–213.
- Kendall, M. G. 1945. The treatment of ties in rank problems. *Biometrika* **33**: 239–251.
- Kendall, M., and A. Stuart. 1979. *The Advanced Theory of Statistics*, Vol. 2; *Inference and Relationship*, 4th ed. New York: Macmillan.
- Kenward, M. G., and B. Jones. 1991. The analysis of categorical data from cross-over trials using a latent variable model. *Stat. Med.* **10**: 1607–1619.
- Kenward, M. G., and B. Jones. 1994. The analysis of binary and categorical data from crossover trials. *Stat. Methods Med. Res.* **3**: 325–344.
- Kenward, M. G., E. Lesaffre, and G. Molenberghs. 1994. An application of maximum likelihood and estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics* **50**: 945–953.
- Khamis, H. J. 1983. Log-linear model analysis of the semi-symmetric intraclass contingency table. *Commun. Stat. A* **12**: 2723–2752.
- Khamis, H. J. 2011. *The Association Graph and the Multigraph for Loglinear Models*. Thousand Oaks, CA: Sage, Publications.
- Khuri A., B. Mukherjee, B. Sinha, and M. Ghosh. 2006. Design issues for generalized linear models: A review. *Stat. Sci.* **21**: 376–399.
- Kim, D., and A. Agresti. 1995. Improved exact inference about conditional association in three-way contingency tables. *J. Am. Stat. Assoc.* **90**: 632–639.
- Kim, D., and A. Agresti. 1997. Nearly exact tests of conditional independence and marginal homogeneity for sparse contingency tables. *Comput. Stat. Data An.* **24**: 89–104.
- King, G. 1997. *A Solution to the Ecological Inference Problem*. Princeton, NJ: Princeton University Press.
- King, G., and L. Zeng. 2001. Logistic regression in rare events data. *Polit. Anal.* **9**: 137–163.
- King, R., and S. P. Brooks. 2001. Prior induction in log-linear models for general contingency table analysis. *Ann. Stat.* **29**: 715–747.
- Klingenberg, B. 2008. Regression models for binary time series with gaps. *Comput. Stat. Data An.* **52**: 4076–4090.
- Kneib, T., and L. Fahrmeir. 2006. Structured additive regression for categorical space-time data: A mixed model approach. *Biometrics* **62**: 109–118.
- Knuiman, M. W., and T. P. Speed. 1988. Incorporating prior information into the analysis of contingency tables. *Biometrics* **44**: 1061–1071.
- Koch, G. G., and V. P. Bhapkar. 1982. Chi-square tests. In *Encyclopedia of Statistical Sciences*, Vol. 1. Hoboken, NJ: Wiley, pp. 442–457.
- Koch, G. G., D. H. Freeman, and J. L. Freeman. 1975. Strategies in the multivariate analysis of data from complex surveys. *Int. Stat. Rev.* **43**: 59–78.
- Koch, G. G., J. R. Landis, J. L. Freeman, D. H. Freeman, and R. G. Lehnen. 1977. A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* **33**: 133–158.
- Koch, G. G., I. A. Arnara, G. W. Davis, and D. B. Gillings. 1982. A review of some statistical methods for covariance analysis of categorical data. *Biometrics* **38**: 563–595.
- Koch, G. G., P. B. Imrey, J. M. Singer, S. S. Atkinson, and M. E. Stokes. 1985. *Lecture Notes for Analysis of Categorical Data*. Montreal: Les Presses de L’Université de Montréal.
- Koehler, K. 1986. Goodness-of-fit tests for log-linear models in sparse contingency tables. *J. Am. Stat. Assoc.* **81**: 483–493.

- Koehler, K. 2005. Chi-square tests. In *Encyclopedia Biostatistics*, 2nd ed. Chichester, UK: Wiley, pp. 765–778.
- Koehler, K., and K. Larntz. 1980. An empirical investigation of goodness-of-fit statistics for sparse multinomials. *J. Am. Stat. Assoc.* **75**: 336–344.
- Koehler, K., and J. Wilson. 1986. Chi-square tests for comparing vectors of proportions for several cluster samples. *Commun. Stat. A* **15**: 2977–2990.
- Koopman, P. A. R. 1984. Confidence limits for the ratio of two binomial proportions. *Biometrics* **40**: 513–517.
- Kosmidis, I., and D. Firth. 2011. Multinomial logit bias reduction via the Poisson log-linear model. *Biometrika* **98**: 755–759.
- Kraemer, H. C. 1979. Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika* **44**: 461–472.
- Kraemer, H. C., V. S. Periyakoil, and A. Noda. 2002. Kappa coefficients in medical research. *Stat. Med.* **21**: 2109–2129.
- Krampe, A., M. Kateri, and S. Kuhnt. 2011. Asymmetry models for square contingency tables: Exact tests via algebraic statistics. *Stat. & Computing* **21**: 55–67.
- Kreiner, S. 1987. Analysis of multidimensional contingency tables by exact conditional tests: Techniques and strategies. *Scand. J. Stat.* **14**: 97–112.
- Kruskal, W. H. 1958. Ordinal measures of association. *J. Am. Stat. Assoc.* **53**: 814–861.
- Ku, H. H., R. N. Varner, and S. Kullback. 1971. On the analysis of multidimensional contingency tables. *J. Am. Stat. Assoc.* **66**: 55–64.
- Kuha, J., and D. Firth. 2011. On the index of dissimilarity for lack of fit in log-linear and log-multiplicative models. *Comput. Stat. Data An.* **55**: 375–388.
- Kuha, J., and C. Skinner. 1997. Categorical data analysis and misclassification. In *Survey Measurement and Process Quality*, eds. L. Lyberg et al. Hoboken, NJ: Wiley, pp. 633–670.
- Kuha, J., C. Skinner, and J. Palmgren. 2005. Misclassification error. In *Encyclopedia Biostatistics*, 2nd ed. Chichester, UK: Wiley, pp. 3257–3264.
- Kullback, S. 1959. *Information Theory and Statistics*. Hoboken, NJ: Wiley.
- Kullback, S., M. Kupperman, and H. H. Ku. 1962. Tests for contingency tables and Markov chains. *Technometrics* **4**: 573–608.
- Kupper, L. L., and J. K. Haseman. 1978. The use of a correlated binomial model for the analysis of certain toxicological experiments. *Biometrics* **34**: 69–76.
- Kupper, L. L., C. Portier, M. D. Hogan, and E. Yamamoto. 1986. The impact of litter effects on dose-response modeling in teratology. *Biometrics* **42**: 85–98.
- Läärä, E., and J. N. S. Matthews. 1985. The equivalence of two models for ordinal data. *Biometrika* **72**: 206–207.
- Lachin, J. M. 1977. Sample-size determinations for $r \times c$ comparative trials. *Biometrics* **33**: 315–324.
- Laird, N. M. 1978. Empirical Bayes methods for two-way contingency tables. *Biometrika* **65**: 581–590.
- Laird, N. M. 2005. EM algorithm. In *Encyclopedia of Biostatistics*, 2nd ed. Chichester, UK: Wiley, pp. 1618–1631.
- Laird, N. M., and D. Olivier. 1981. Covariance analysis of censored survival data using log-linear analysis techniques. *J. Am. Stat. Assoc.* **76**: 231–240.
- Lancaster, H. O. 1949a. The derivation and partition of χ^2 in certain discrete distributions. *Biometrika* **36**: 117–129.
- Lancaster, H. O. 1949b. The combination of probabilities arising from data in discrete distributions.

- Biometrika* **36**: 370–382.
- Lancaster, H. O. 1961. Complex contingency tables treated by partition of χ^2 . *J. R. Stat. Soc. B* **13**: 242–249.
- Lancaster, H. O. 1951. Significance tests in discrete distributions. *J. Am. Stat. Assoc.* **56**: 223–234.
- Lancaster, H. O. 1969. *The Chi-Squared Distribution*. Hoboken, NJ: Wiley.
- Lancaster, H. O., and M. A. Hamdan. 1964. Estimation of the correlation coefficient in contingency tables with possible nonmetrical characters. *Psychometrika* **29**: 383–391.
- Landis, J. R., and G. G. Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* **33**: 363–374.
- Landis, J. R., E. R. Heyman, and G. G. Koch. 1978. Average partial association in three-way contingency tables: A review and discussion of alternative tests. *Int. Stat. Rev.* **46**: 237–254.
- Landis, J. R., T. J. Sharp, S. J. Kuritz, and G. G. Koch. 2005. Mantel–Haenszel methods. In *Encyclopedia of Biostatistics*, 2nd ed. Chichester, UK: Wiley, pp. 2937–2950.
- Landis, J. R., T. S. King, J. W. Choi, V. M. Chinchilli, and G. G. Koch. 2011. Measures of agreement and concordance with clinical research applications. *Stat. Biopharm. Res.* **3**: 185–209.
- Landwehr, J. M., D. Pregibon, and A. C. Shoemaker. 1984. Graphical methods for assessing logistic regression models. *J. Am. Stat. Assoc.* **79**: 61–71.
- Lang, J. B. 1992. Obtaining the observed information matrix for the Poisson log linear model with incomplete data. *Biometrika* **79**: 405–407.
- Lang, J. B. 1996a. Maximum likelihood methods for a generalized class of log-linear models. *Ann. Stat.* **24**: 726–752.
- Lang, J. B. 1996b. On the partitioning of goodness-of-fit statistics for multivariate categorical response models. *J. Am. Stat. Assoc.* **91**: 1017–1023.
- Lang, J. B. 1996c. On the comparison of multinomial and Poisson log-linear models. *J. R. Stat. Soc. B* **58**: 253–266.
- Lang, J. B. 1999. Bayesian ordinal and binary regression models with a parametric family of mixture links. *Comput. Stat. Data An.* **31**: 59–87.
- Lang, J. B. 2004. Multinomial-Poisson homogeneous models for contingency tables. *Ann. Stat.* **32**: 340–383.
- Lang, J. B. 2005. Homogeneous linear predictor models for contingency tables. *J. Am. Stat. Assoc.* **100**: 121–134.
- Lang, J. B. 2008. Score and profile likelihood confidence intervals for contingency table parameters. *Stat. Med.* **27**: 5975–5990.
- Lang, J. B., and A. Agresti. 1994. Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *J. Am. Stat. Assoc.* **89**: 625–632.
- Lang, J. B., J. W. McDonald, and P. W. F. Smith. 1999. Association-marginal modeling of multivariate categorical responses: A maximum likelihood approach. *J. Am. Stat. Assoc.* **94**: 1161–1171.
- Laplace, P. S. 1812. *Théorie Analytique des Probabilités*. Paris: Courcier.
- Lapointe, F.-J., and P. Legendre. 1994. A classification of pure malt Scotch whiskies. *Appl. Stat.* **43**: 237–257.
- Larntz, K. 1978. Small-sample comparison of exact levels for chi-squared goodness-of-fit statistics. *J. Am. Stat. Assoc.* **73**: 253–263.
- Larsen, K., J. H. Petersen, E. Budtz-Jørgensen, and L. Endahl. 2000. Interpreting parameters in the logistic regression model with random effects. *Biometrics* **56**: 909–914.
- Larson, M. G. 1984. Covariate analysis of competing-risks data with log-linear models. *Biometrics* **40**: 459–469.

- Lauritzen, S. L. 1996. *Graphical Models*. New York: Oxford University Press.
- Lauritzen, S. L., and N. Wermuth. 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Stat.* **17**: 31–57.
- LaVange, L. M., G. G. Koch, and T. A. Schwartz. 2001. Applying sample survey methods to clinical trials data. *Stat. Med.* **20**: 2609–2623.
- Lawless, J. F. 1987. Negative binomial and mixed Poisson regression. *Can. J. Stat.* **15**: 209–225.
- Lazarsfeld, P. F., and N. W. Henry. 1968. *Latent Structure Analysis*. Boston: Houghton Mifflin.
- le Cessie, S., and J. van Houwelingen. 1992. Ridge estimators in logistic regression. *Appl. Stat.* **41**: 191–201.
- Lee, A., and M. Silvapulle. 1988. Ridge estimation in logistic regression. *Commun. Stat. Simul. Comput.* **17**: 1231–1257.
- Lee, S. K. 1977. On the asymptotic variances of terms in loglinear models of multidimensional contingency tables. *J. Am. Stat. Assoc.* **72**: 412–419.
- Lee, Y., and J. A. Nelder. 1996. Hierarchical generalized linear models. *J. R. Stat. Soc. B* **58**: 619–678.
- Lee, Y., J. A. Nelder, and Y. Pawitan. 2006. *Generalized Linear Models with Random Effects: Unified Analysis via H-Likelihood*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Lefkopoulou, M., D. Moore, and L. Ryan. 1989. The analysis of multiple correlated binary outcomes: Application to rodent teratology experiments. *J. Am. Stat. Assoc.* **84**: 810–815.
- Lehmann, E. L. 1966. Some concepts of dependence. *Ann. Math. Stat.* **37**: 1137–1153.
- Lehmann, E. L., and J. P. Romano. 2005. *Testing Statistical Hypotheses*, 3rd ed. New York: Springer.
- Leonard, T. 1972. Bayesian methods for binomial data. *Biometrika* **59**: 581–589.
- Leonard, T. 1975. Bayesian estimation methods for two-way contingency tables. *J. R. Stat. Soc. B* **37**: 23–37.
- Leonard, T. 1999. *A Course in Categorical Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Leonard, T., and J. S. J. Hsu. 1994. The Bayesian analysis of categorical data: A selective review. In *Aspects of Uncertainty: A Tribute to D. V. Lindley*, eds. P. R. Freeman and A. F. M. Smith. Hoboken, NJ: Wiley, pp. 283–310.
- Lesaffre, E., and A. Albert. 1989. Multiple-group logistic regression diagnostics. *Appl. Stat.* **38**: 425–440.
- Lesaffre, E., and G. Molenberghs. 1991. Multivariate probit analysis: A neglected procedure in medical statistics. *Stat. Med.* **10**: 1391–1403.
- Lesaffre, E., and B. Spiessens. 2001. On the effect of the number of quadrature points in a logistic random effects model: An example. *Appl. Stat.* **50**: 325–335.
- Lesaffre, E., D. Rizopoulos, and R. Tsonaka. 2007. The logistic transform for bounded outcome scores. *Biostatistics* **8**: 72–85.
- Lewis, T., I. W. Saunders, and M. Westcott. 1984. The moments of the Pearson chi-squared statistic and the minimum expected value in two-way tables. *Biometrika* **71**: 515–522.
- Li, D., and D. V. Conti. 2009. Detecting gene–environment interactions using a combined case-only and case–control approach. *Am. J. Epidemiol.* **169**: 497–504.
- Li, P. 2010. Robust logitboost and adaptive base class (ABC) logitboost. Conference on Uncertainty in Artificial Intelligence.
- Liang, K. Y. 1984. The asymptotic efficiency of conditional likelihood methods. *Biometrika* **71**: 305–313.
- Liang, K. Y., and J. Hanfelt. 1994. On the use of the quasi-likelihood method in teratological

- experiments. *Biometrics* **50**: 872–880.
- Liang, K. Y., and P. McCullagh. 1993. Case studies in binary dispersion. *Biometrics* **49**: 623–630.
- Liang, K. Y., and S. G. Self. 1985. Tests for homogeneity of odds ratios when the data are sparse. *Biometrika* **72**: 353–358.
- Liang, K. Y., and S. L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* **73**: 13–22.
- Liang, K. Y., and S. L. Zeger. 1988. On the use of concordant pairs in matched case-control studies. *Biometrics* **44**: 1145–1156.
- Liang, K. Y., and S. L. Zeger. 1989. A class of logistic regression models for multivariate binary time series. *J. Am. Stat. Assoc.* **84**: 447–451.
- Liang, K. Y., and S. L. Zeger. 1995. Inference based on estimating functions in the presence of nuisance parameters. *Stat. Sci.* **10**: 158–173.
- Liang, K. Y., S. L. Zeger, and B. Qaqish. 1992. Multivariate regression analyses for categorical data. *J. R. Stat. Soc. B* **54**: 3–24.
- Liao, J. G., and D. McGee. 2003. Adjusted coefficients of determination for logistic regression. *Am. Stat.* **57**: 161–165.
- Lin, C.-Y., and M.-C. Yang. 2006. Improved exact confidence intervals for the odds ratio in two independent binomial samples. *Biomet. J* **48**: 1008–1019.
- Lin, C.-Y., and M.-C. Yang. 2009. Improved *p*-value tests for comparing two independent binomial proportions. *Commun. Statist. Simul. Comput.* **38**: 78–91.
- Lin, H., Z. Guo, P. N. Peduzzi, T. M. Gill, and H. G. Allore. 2008. A semiparametric transition model with latent traits for longitudinal multistate data. *Biometrics* **64**: 1032–1042.
- Lin, X., T. Cai, and A. Schwartzman. 2010. Statistical methods for analysis of high-dimensional data with applications in biosciences. Short course presented at ENAR meeting in New Orleans.
- Lindley, D. V. 1964. The Bayesian analysis of contingency tables. *Ann. Math. Stat.* **35**: 1622–1643.
- Lindsay, B., C. Clogg, and J. Grego. 1991. Semi-parametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *J. Am. Stat. Assoc.* **86**: 96–107.
- Lindsey, J. K., and P. M. E. Altham. 1998. Analysis of the human sex ratio by using overdispersion models. *Appl. Stat.* **47**: 149–157.
- Lindsey, J. K., and P. Lambert. 1999. On the appropriateness of marginal models for repeated measurements in clinical trials. *Stat. Med.* **17**: 447–469.
- Lipsitz, S. 1992. Methods for estimating the parameters of a linear model for ordered categorical data. *Biometrics* **48**: 271–281.
- Lipsitz, S. R., and G. Fitzmaurice. 1996. The score test for independence in $R \times C$ contingency tables with missing data. *Biometrics* **52**: 751–762.
- Lipsitz, S., N. Laird, and D. Harrington. 1990. Finding the design matrix for the marginal homogeneity model. *Biometrika* **77**: 353–358.
- Lipsitz, S., N. Laird, and D. Harrington. 1991. Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika* **78**: 153–160.
- Lipsitz, S. R., K. Kim, and L. Zhao. 1994. Analysis of repeated categorical data using generalized estimating equations. *Stat. Med.* **13**: 1149–1163.
- Little, R. J. 1989. Testing the equality of two independent binomial proportions. *Am. Stat.* **43**: 283–288.
- Little, R. J. 2005. Missing data. In *Encyclopedia of Biostatistics*, 2nd ed. Chichester, UK: Wiley, pp. 3272–3285.
- Little, R. J., and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*, 2nd ed. Hoboken, NJ:

Wiley.

- Little, R. J. A., and M.-M. Wu. 1991. Models for contingency tables with known margins when target and sampled populations differ. *J. Am. Stat. Assoc.* **86**: 87–95.
- Liu, D., D. Ghosh, and X. Lin. 2008. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics* **9**: 292.
- Liu, J. S., A. F. Neuwald, and C. E. Lawrence. 1995. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.* **90**: 1156–1170.
- Liu, Q., and D. A. Pierce. 1993. Heterogeneity in Mantel–Haenszel-type models. *Biometrika* **80**: 543–556.
- Liu, Q., and D. A. Pierce. 1994. A note on Gauss–Hermite quadrature. *Biometrika* **81**: 624–629.
- Lloyd, C. J. 1988a. Some issues arising from the analysis of 2×2 contingency tables. *Austral. J. Stat.* **30**: 35–46.
- Lloyd, C. J. 1988b. Doubling the one-sided P -value in testing independence in 2×2 tables against a two-sided alternative. *Stat. Med.* **7**: 1297–1306.
- Lloyd, C. J. 1999. *Statistical Analysis of Categorical Data*. Hoboken, NJ: Wiley.
- Lloyd, C. J. 2008. A new exact and more powerful unconditional test of no treatment effect from binary matched pairs. *Biometrics* **64**: 716–723.
- Loh, W.-Y. 2002. Regression trees with unbiased variable selection and interaction detection. *Stat. Sin.* **12**: 361–386.
- Loh, W.-Y., and Y.-S. Shih. 1997. Split selection methods for classification trees. *Stat. Sin.* **7**: 815–840.
- Longford, N. T. 1993. *Random Coefficient Models*. New York: Oxford University Press.
- Loughin, T. M., and P. N. Scherer. 1998. Testing for association in contingency tables with multiple column responses. *Biometrics* **54**: 630–637.
- Louis, T. A. 1982. Finding the observed information matrix when using the EM algorithm. *J. R. Stat. Soc. B* **44**: 226–233.
- Lovison, G. 2005. On Rao score and Pearson X^2 statistics in generalized linear models. *Stat. Pap.* **46**: 555–574.
- Luce, R. D. 1959. *Individual Choice Behavior*. Hoboken, NJ: Wiley.
- Lyles, R. H., H.-M. Lin, and J. M. Williamson. 2006. A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses. *Stat. Med.* **26**: 1632–1648.
- Madansky, A. 1963. Tests of homogeneity for correlated samples. *J. Am. Stat. Assoc.* **58**: 97–119.
- Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge, UK: Cambridge University Press.
- Madigan, D., and A. E. Raftery. 1994. Model selection and accounting for model uncertainty in graphical models using Occam’s window. *J. Am. Stat. Assoc.* **89**: 1535–1546.
- Madigan, D., and J. York. 1995. Bayesian graphical models for discrete data. *Int. Stat. Rev.* **63**: 215–232.
- Magidson, J. 2010. Correlated component regression: A prediction/classification methodology for possibly many features. *JSM Proc. Am. Stat. Assoc., Sec. on Statistical Learning and Data Mining*, 4372–4386.
- Magidson, J., and J. K. Vermunt. 2004. Latent class models. In *The Sage Handbook of Quantitative Methodology for the Social Sciences*, ed. D. Kaplan. Thousand Oaks, CA: Sage Publications, Chap. 10, pp. 175–198.
- Mai, Q., H. Zou, and M. Yuan. 2012. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika* **99**: 29–42.

- Maiste, P. J., and B. S. Weir. 2004. Optimal testing strategies for large, sparse multinomial models. *Comput. Stat. Data An.* **46**: 605–620.
- Mantel, N. 1963. Chi-square tests with one degree of freedom: Extensions of the Mantel–Haenszel procedure. *J. Am. Stat. Assoc.* **58**: 690–700.
- Mantel, N. 1966. Models for complex contingency tables and polychotomous dosage response curves. *Biometrics* **22**: 83–95.
- Mantel, N. 1973. Synthetic retrospective studies and related topics. *Biometrics* **29**: 479–486.
- Mantel, N. 1985. Maximum likelihood vs. minimum chi-square. *Biometrics* **41**: 777–781.
- Mantel, N. 1987a. Understanding Wald’s test for exponential families. *Am. Stat.* **41**: 147–148.
- Mantel, N. 1987b. Exact tests for 2×2 contingency tables (Letter). *Am. Stat.* **41**: 159.
- Mantel, N., and D. P. Byar. 1978. Marginal homogeneity, symmetry and independence. *Commun. Stat. A* **7**: 953–976.
- Mantel, N., and J. L. Fleiss. 1980. Minimum expected cell size requirements for the Mantel-Haenszel one-degree-of-freedom chi-square test and a related rapid procedure. *Am. J. Epidemiol.* **112**: 129–134.
- Mantel, N., and W. Haenszel. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22**: 719–748.
- Marchetti, G. M., and M. Lupparelli. 2010. Chain graph models of multivariate regression type for categorical data. *Bernoulli* **17**: 827–841.
- Martín Andrés, A., and A. Silva Mato, 1994. Choosing the optimal unconditional test for comparing two independent proportions. *Comput. Stat. Data An.* **17**: 555–574.
- Massam, H., J. Liu, and A. Dobra. 2009. A conjugate prior for discrete hierarchical log-linear models. *Ann. Stat.* **37**: 3431–3467.
- Matthews, J. N. S., and K. P. Morris. 1995. An application of Bradley–Terry-type models to the measurement of pain. *Appl. Stat.* **44**: 243–255.
- McCullagh, P. 1980. Regression models for ordinal data. *J. R. Stat. Soc. B* **42**: 109–142.
- McCullagh, P. 1982. Some applications of quasisymmetry. *Biometrika* **69**: 303–308.
- McCullagh, P. 1983. Quasi-likelihood functions. *Ann. Stat.* **11**: 59–67.
- McCullagh, P. 1986. The conditional distribution of goodness-of-fit statistics for discrete data. *J. Am. Stat. Assoc.* **81**: 104–107.
- McCullagh, P. 2008. Sampling bias and logistic models. *J. R. Stat. Soc. B* **70**: 643–677.
- McCullagh, P., and J. A. Nelder. 1983; 2nd ed., 1989. *Generalized Linear Models*. London: Chapman & Hall.
- McCulloch, C. E. 2000. Generalized linear models. *J. Am. Stat. Assoc.* **95**: 1320–1324.
- McCulloch, C. E., and J. M. Neuhaus. 2011. Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics* **67**: 270–279.
- McCulloch, C. E., S. Searle, and J. M. Neuhaus. 2008. *Generalized, Linear, and Mixed Models*. Hoboken, NJ: Wiley.
- McCulloch, R. E., N. G. Polson, and P. E. Rossi. 2000. A Bayesian analysis of the multinomial probit model with fully identified parameters. *J. Econometrics* **99**: 173–193.
- McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, ed. P. Zarembka. New York: Academic Press, pp. 105–142.
- McFadden, D. 1980. Econometric models of probabilistic choice among products. *J. Business* **53**: S13–29.
- McFadden, D., and K. Train. 2000. Mixed MNL models for discrete response. *J. Appl. Econometrics* **15**: 447–470.
- McKelvey, R. D., and W. Zavoina. 1975. A statistical model for the analysis of ordinal level

- dependent variables. *J. Math. Sociol.* **4**: 103–120.
- McLachlan, G. J. 2004. *Discriminant Analysis and Statistical Pattern Recognition*. Hoboken, NJ: Wiley.
- McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**: 153–157.
- Mee, R. W. 1984. Confidence bounds for the difference between two probabilities (letter). *Biometrics* **40**: 1175–1176.
- Meeden, G., C. Geyer, J. Lang, and E. Funo. 1998. The admissibility of the maximum likelihood estimator for decomposable log-linear interaction models for contingency tables. *Commun. Stat. A* **27**: 473–494.
- Mehrotra, D. V., I. S. F. Chan, and R. L. Berger. 2003. A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics* **59**: 441–450.
- Mehta, C. R. 1994. The exact analysis of contingency tables in medical research. *Stat. Methods Med. Res.* **3**: 135–156.
- Mehta, C. R., and N. R. Patel. 1983. A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *J. Am. Stat. Assoc.* **78**: 427–434.
- Mehta, C. R., and N. R. Patel. 1995. Exact logistic regression: Theory and examples. *Stat. Med.* **14**: 2143–2160.
- Mehta, C. R., N. R. Patel, and R. Gray. 1985. Computing an exact confidence interval for the common odds ratio in several 2 by 2 contingency tables. *J. Am. Stat. Assoc.* **80**: 969–973.
- Mehta, C. R., N. R. Patel, and P. Senchaudhuri. 2000. Efficient Monte Carlo methods for conditional logistic regression. *J. Am. Stat. Assoc.* **95**: 99–108.
- Meier, L., S. van de Geer, and P. Bühlmann. 2008. The group lasso for logistic regression. *J. R. Stat. Soc. B* **70**: 53–71.
- Menard, S. 2004. Six approaches to calculating standardized logistic regression coefficients. *Am. Stat.* **58**: 218–223.
- Meng, R. C., and D. G. Chapman. 1966. The power of chi-square tests for contingency tables. *J. Am. Stat. Assoc.* **61**: 965–975.
- Meulman, J. J. 2003. Prediction and classification in nonlinear data analysis: Something old, something new, something borrowed, something blue. *Psychometrika* **68**: 493–517.
- Michailidis, G., and J. de Leeuw. 1998. The Gifi system of descriptive multivariate analysis. *Stat. Sci.* **13**: 307–336.
- Miettinen, O. S. 1969. Individual matching with multiple controls in the case of all-or-none responses. *Biometrics* **25**: 339–355.
- Miettinen, O. S., and M. Nurminen. 1985. Comparative analysis of two rates. *Stat. Med.* **4**: 213–226.
- Miller, J., and J. Franklin. 2002. Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. *Ecolog. Modelling* **157**: 227–247.
- Miller, M. E., C. S. Davis, and J. R. Landis. 1993. The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares. *Biometrics* **49**: 1033–1044.
- Min, Y., and A. Agresti. 2005. Random effects models for repeated measures of zero-inflated count data. *Statist. Modelling* **5**: 1–19.
- Mirkin, B. 2001. Eleven ways to look at the chi-squared coefficient for contingency tables. *Am. Stat.* **55**: 111–120.
- Mitra, S. K. 1958. On the limiting power function of the frequency chi-square test. *Ann. Stat.* **29**: 1221–1233.

- Mittlböck, M., and M. Schemper. 1996. Explained variation for logistic regression. *Stat. Med.* **15**: 1987–1997.
- Molenberghs, G., and E. Goetghebeur. 1997. Simple fitting algorithms for incomplete categorical data. *J. R. Stat. Soc. B* **59**: 401–414.
- Molenberghs, G., and E. Lesaffre. 1994. Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. *J. Am. Stat. Assoc.* **89**: 633–644.
- Molenberghs, G., and E. Lesaffre. 1999. Marginal modelling of multivariate categorical data. *Stat. Med.* **18**: 2237–2255.
- Molenberghs, G., and G. Verbeke. 2005. *Models for Discrete Longitudinal Data*. New York: Springer.
- Molenberghs, G., and G. Verbeke. 2007. Likelihood ratio, score and Wald tests in a constrained parameter space. *Am. Stat.* **61**: 22–27.
- Molenberghs, G., G. Verbeke, C. G. B. Demétrio, and A. M. C. Vieira. 2010. A family of generalized linear models for repeated measures with normal and conjugate random effects. *Stat. Sci.* **25**: 325–347.
- Moore, D. F. 1986a. Asymptotic properties of moment estimates for overdispersed counts and proportions. *Biometrika* **35**: 583–588.
- Moore, D. F., and A. Tsiatis. 1991. Robust estimation of the variance in moment methods for extra-binomial and extra-Poisson variation. *Biometrics* **47**: 383–401.
- Moore, D. S. 1986b. Tests of chi-squared type. In *Goodness-of-Fit Techniques*, eds. R. D'Agostino and M. A. Stephens. New York: Marcel Dekker, pp. 63–95.
- Morris, C. 1975. Central limit theorems for multinomial sums. *Ann. Stat.* **3**: 165–188.
- Mosimann, J. E. 1962. On the compound multinomial distribution, the multivariate β -distribution and correlations among proportions. *Biometrika* **49**: 65–82.
- Mosteller, F. 1951. Remarks on the method of paired comparisons I: The least-squares solution assuming equal standard deviations and equal correlations. *Psychometrika* **16**: 3–9.
- Mosteller, F. 1968. Association and estimation in contingency tables. *J. Am. Stat. Assoc.* **63**: 1–28.
- Muenz, L. R., and L. V. Rubinstein. 1985. Markov models for covariate dependence of binary sequences. *Biometrics* **41**: 91–101.
- Mukherjee, B., and N. Chatterjee. 2008. Exploiting gene–environment independence for analysis of case–control studies: An empirical Bayes-type shrinkage estimator to trade off between bias and efficiency. *Biometrics* **64**: 685–694.
- Mukherjee, B., and I. Liu. 2008. A characterization of bias for fitting multivariate generalized linear models under choice-based sampling. *J. Multiv. Anal.* **100**: 459–472.
- Müller, P., and K. Roeder. 1997. A Bayesian semiparametric model for case–control studies with errors in variables. *Biometrika* **84**: 523–537.
- Nair, V. N. 1987. Chi-squared-type tests for ordered alternatives in contingency tables. *J. Am. Stat. Assoc.* **82**: 283–291.
- Natarajan, R., and C. McCulloch. 1995. A note on the existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika* **82**: 639–643.
- Natarajan, R., C. McCulloch, and N. M. Kiefer. 2000. A Monte Carlo EM method for estimating multinomial probit models. *Comput. Stat. Data An.* **34**: 33–50.
- Natarajan, S., S. R. Lipsitz, and R. Gonin. 2008. Variance estimation in complex survey sampling for generalized linear models. *Appl. Stat.* **57**: 75–87.
- Natarajan, S., S. Lipsitz, G. M. Fitzmaurice, D. Sinha, J. G. Ibrahim, J. Haas, and W. Gellad. 2012. An extension of the Wilcoxon rank-sum test for complex sample survey data. *Appl. Stat.* **61**: 653–664.

- Nelder, J., and D. Pregibon. 1987. An extended quasi-likelihood function. *Biometrika* **74**: 221–232.
- Nelder, J., and R. W. M. Wedderburn. 1972. Generalized linear models. *J. R. Stat. Soc. A* **135**: 370–384.
- Neuhaus, J. M. 1992. Statistical methods for longitudinal and clustered designs with binary responses. *Stat. Methods Medic. Res.* **1**: 249–273.
- Neuhaus, J. M., and N. P. Jewell. 1990a. Some comments on Rosner's multiple logistic model for clustered data. *Biometrics* **46**: 523–534.
- Neuhaus, J. M., and N. P. Jewell. 1990b. The effect of retrospective sampling on binary regression models for clustered data. *Biometrics* **46**: 977–990.
- Neuhaus, J. M., and M. L. Lesperance. 1996. Estimation efficiency in a binary mixed-effects model setting. *Biometrika* **83**: 441–446.
- Neuhaus, J. M., J. D. Kalbfleisch, and W. W. Hauck. 1991. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Int. Stat. Rev.* **59**: 25–35.
- Neuhaus, J. M., W. W. Hauck, and J. D. Kalbfleisch. 1992. The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79**: 755–762.
- Neuhaus, J. M., J. D. Kalbfleisch, and W. W. Hauck. 1994. Conditions for consistent estimation in mixed-effects models for binary matched-pairs data. *Can. J. Stat.* **22**: 139–148.
- Newcombe, R. 1998a. Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Stat. Med.* **17**: 857–872.
- Newcombe, R. 1998b. Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Stat. Med.* **17**: 873–890.
- Newcombe, R. 2001. Logit confidence intervals and the inverse sinh transformation. *Am. Stat.* **55**: 200–202.
- Neyman, J. 1935. On the problem of confidence limits. *Ann. Math. Stat.* **6**: 111–116.
- Neyman, J. 1949. Contributions to the theory of the χ^2 test. In *Proceedings First Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Berkeley, CA: University of California Press, pp. 239–273.
- Noe, D. A., I. M. Nelson, S. Mehdizadeh, and A. J. Bailer. 2009. Will they stay or will they go? Predicting disenrollment from home care services to nursing homes using classification trees. *Chance* **22**(4): 58–62.
- Ntzoufras, I., J. J. Forster, and P. Dellaportas. 2000. Stochastic search variable selection for log-linear models. *J. Stat. Comput. Simul.* **68**: 23–37.
- Nurminen, M. 1986. Confidence intervals for the ratio and difference of two binomial proportions. *Biometrics* **42**: 675–676.
- Nurminen, M., and P. Mutanen. 1987. Exact Bayesian analysis of two proportions. *Scand. J. Stat.* **14**: 67–77.
- O'Brien, P. C. 1988. Comparing two samples: Extensions of the t , rank-sum, and log-rank tests. *J. Am. Stat. Assoc.* **83**: 52–61.
- O'Brien, R. G. 1986. Using the SAS system to perform power analyses for log-linear models. In *Proceedings 11th Annual SAS Users Group Conference*. Cary, NC: SAS Institute, pp. 778–784.
- O'Brien, S. M., and D. B. Dunson. 2004. Bayesian multivariate logistic regression. *Biometrics* **60**: 739–746.
- Ochi, Y., and R. Prentice. 1984. Likelihood inference in a correlated probit regression model. *Biometrika* **71**: 531–543.
- O'Gorman, T. W., and R. F. Woolson. 1988. Analysis of ordered categorical data using the SAS system. In *Proceedings 13th Annual SAS Users Group Conference*. Cary, NC: SAS Institute, pp. 957–963.

- O'Hagan, A., and J. Forster. 2004. *Kendall's Advanced Theory of Statistics: Bayesian Inference*, London: Arnold.
- Olive, D., and D. Hawkins. 2005. Variable selection for 1D regression models. *Technometrics* **47**: 43–50.
- Osius, G., and D. Rojek. 1992. Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *J. Am. Stat. Assoc.* **87**: 1145–1152.
- Owen, A. B. 2007. Infinitely imbalanced logistic regression. *J. Machine Learn. Res.* **8**: 761–773.
- Paige, R. L., P. L. Chapman, and R. W. Butler. 2011. Small sample LD50 confidence intervals using saddlepoint approximations. *J. Am. Stat. Assoc.* **106**: 334–344.
- Paik, M. 1985. A graphic representation of a three-way contingency table: Simpson's paradox and correlation. *Am. Stat.* **39**: 53–54.
- Palmgren, J. 1981. The Fisher information matrix for log-linear models arguing conditionally in the observed explanatory variables. *Biometrika* **68**: 563–566.
- Palmgren, J., and A. Ekholm. 1987. Exponential family non-linear models for categorical data with errors of observation. *Appl. Stochastic Models Data Anal.* **3**: 111–124.
- Park, M. Y., and T. Hastie. 2008. Penalized logistic regression for detecting gene interactions. *Biostatistics* **9**: 30–50.
- Park, T., and M. B. Brown. 1994. Models for categorical data with nonignorable nonresponse. *J. Am. Stat. Assoc.* **89**: 44–52.
- Parsons, N. R., M. L. Costa, J. Achten, and N. Stallard. 2009. Repeated measures proportional odds logistic regression analysis of ordinal score data in the statistical software package R. *Comput. Stat. Data An.* **53**: 632–641.
- Parzen, E. 1997. Concrete statistics. In *Statistics of Quality*. New York: Marcel Dekker, pp. 309–332.
- Parzen, M., S. Ghosh, S. Lipsitz, D. Sinha, G. Fitzmaurice, B. Mallick, and J. Ibrahim. 2011. A generalized linear mixed model for longitudinal binary data with a marginal logit link function. *Ann. Appl. Stat.* **5**: 449–467.
- Patefield, W. M. 1982. Exact tests for trends in ordered contingency tables. *Appl. Stat.* **31**: 32–43.
- Patnaik, P. B. 1949. The non-central χ^2 and F-distributions and their applications. *Biometrika* **36**: 202–232.
- Paul, S. R., K. Y. Liang, and S. G. Self. 1989. On testing departure from the binomial and multinomial assumptions. *Biometrics* **45**: 231–236.
- Pavlides, M. G., and M. D. Perlman. 2009. How likely is Simpson's paradox? *Am. Stat.* **63**: 226–233.
- Pearson, E. S. 1947. The choice of a statistical test illustrated on the interpretation of data classified in 2×2 tables. *Biometrika* **34**: 139–167.
- Pearson, K. 1900. On a criterion that a given system of deviations from the probable in the case of a correlated in system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag. Ser. 5* **50**: 157–175. (Reprinted in *Karl Pearson's Early Statistical Papers*, ed. E. S. Pearson. Cambridge, UK: Cambridge University Press, 1948.)
- Pearson, K. 1904. Mathematical contributions to the theory of evolution XIII: On the theory of contingency and its relation to association and normal correlation. *Drapers' Co. Research Memoirs, Biometric Series*, no. 1. (Reprinted in *Karl Pearson's Early Papers*, ed. E. S. Pearson, Cambridge, UK: Cambridge University Press, 1948.)
- Pearson, K. 1909. On a new method of determining correlation between a measured character A, and a character B, of which only the percentage of cases wherein B exceeds (or falls short of) a given intensity is recorded for each grade of A. *Biometrika* **7**: 96–105.
- Pearson, K. 1913. On the probable error of a correlation coefficient as found from a fourfold table.

Biometrika **9**: 22–27.

- Pearson, K. 1917. On the general theory of multiple contingency with special reference to partial contingency. *Biometrika* **11**: 145–158.
- Pearson, K. 1922. On the χ^2 test of goodness of fit. *Biometrika*, **14**: 186–191.
- Pearson, K., and D. Heron. 1913. On theories of association. *Biometrika* **9**: 159–315.
- Peduzzi, P., J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein. 1996. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* **49**: 1373–1379.
- Pendergast, J. F., S. J. Gange, M. A. Newton, M. J. Lindstrom, M. Palta, and M. R. Fisher. 1996. A survey of methods for analyzing clustered binary response data. *Int. Stat. Rev.* **64**: 89–118.
- Pepe, M. S. 2004. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford: Oxford University Press.
- Peterson, B., and F. E. Harrell, Jr. 1990. Partial proportional odds models for ordinal response variables. *Appl. Stat.* **39**: 205–217.
- Piccarreta, R. 2008. Classification trees for ordinal variables. *Comput. Stat.* **23**: 407–427.
- Piegorsch, W. W., C. R. Weinberg, and J. A. Taylor. 1994. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat. Med.* **13**: 153–162.
- Pierce, D. A., and D. Peters. 1992. Practical use of higher order asymptotics for multiparameter exponential families. *J. R. Stat. Soc. B* **54**: 701–725.
- Pierce, D. A., and D. Peters. 1999. Improving on exact tests by approximate conditioning. *Biometrika* **86**: 265–277.
- Pierce, D. A., and B. R. Sands. 1975. Extra-Bernoulli variation in regression of binary data. Tech. Report 46, Statistics Dept., Oregon State University, Corvallis, OR.
- Pierce, D. A., and D. W. Schafer. 1986. Residuals in generalized linear models. *J. Am. Stat. Assoc.* **81**: 977–983.
- Pires, A. M., and J. A. Branco. 2010. A statistical model to explain the Mendel–Fisher controversy. *Stat. Sci.* **25**: 545–565.
- Plackett, R. L. 1962. A note on interactions in contingency tables. *J. R. Stat. Soc. B* **24**: 162–166.
- Plackett, R. L. 1977. The marginal totals of a 2×2 table. *Biometrika* **64**: 37–42.
- Plackett, R. L. 1983. Karl Pearson and the chi-squared test. *Int. Stat. Rev.* **51**: 59–72.
- Pledger, S., K. H. Pollock, and J. L. Norris. 2010. Open capture–recapture models with heterogeneity: II. Jolly–Seber model. *Biometrics* **66**: 883–890.
- Poisson, S.-D. 1837. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités*. Paris: Bachelier.
- Pratt, J. W. 1981. Concavity of the log likelihood. *J. Am. Stat. Assoc.* **76**: 103–106.
- Pregibon, D. 1980. Goodness of link tests for generalized linear models. *Appl. Stat.* **29**: 15–24.
- Pregibon, D. 1981. Logistic regression diagnostics. *Ann. Stat.* **9**: 705–724.
- Pregibon, D. 1982. Score tests in GLIM with application. In *Lecture Notes in Statistics, 14: GLIM 82, Proc. International Conference on Generalised Linear Models*, ed. R. Gilchrist. New York: Springer-Verlag, pp. 87–97.
- Prentice, R. 1975. Discrimination among some parametric models. *Biometrika* **62**: 607–614.
- Prentice, R. 1976a. Use of the logistic model in retrospective studies. *Biometrics* **32**: 599–606.
- Prentice, R. 1976b. Generalization of the probit and logit methods for dose response curves. *Biometrics* **32**: 761–768.
- Prentice, R. 1986. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *J. Am. Stat. Assoc.* **81**: 321–327.

- Prentice, R., and N. Breslow. 1978. Retrospective studies and failure time models. *Biometrika* **65**: 153–158.
- Prentice, R., and L. A. Gloeckler. 1978. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* **34**: 57–67.
- Prentice, R., and R. Pyke. 1979. Logistic disease incidence models and case-control studies. *Biometrika* **66**: 403–412.
- Prentice, R., and L. P. Zhao. 1991. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47**: 825–839.
- Presnell, B., and D. D. Boos. 2004. The IOS test for model misspecification. *J. Am. Stat. Assoc.* **99**: 216–227.
- Press, S. J., and S. Wilson. 1978. Choosing between logistic regression and discriminant analysis. *J. Am. Stat. Assoc.* **73**: 699–705.
- Qin, J., and K.-Y. Liang. 2011. Hypothesis testing in a mixture case-control model. *Biometrics* **67**: 182–193.
- Quine, M. P., and E. Seneta. 1987. Bortkiewicz's data and the law of small numbers. *Int. Stat. Rev.* **5**: 173–181.
- Quintana, F. A. 1998. Nonparametric Bayesian analysis for assessing homogeneity in $k \times l$ contingency tables with fixed right margin totals. *J. Am. Stat. Assoc.* **93**: 1140–1149.
- Rabbee, N., B. A. Coull, C. Mehta, N. Patel, and P. Senchaudhuri. 2003. Power and sample size for ordered categorical data. *Stat. Methods Med. Res.* **12**: 73–84.
- Rabe-Hesketh, S., and A. Skrondal. 2001. Parameterisation of multivariate random effects models for categorical data. *Biometrics* **57**: 1256–1263.
- Rabe-Hesketh, S., A. Skrondal, and A. Pickles. 2005. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *J. Econometrics* **128**: 301–323.
- Racine, A., A. P. Grieve, H. Fluhler, and A. F. M. Smith. 1986. Bayesian methods in practice: Experiences in the pharmaceutical industry. *Appl. Stat.* **35**: 93–150.
- Raftery, A. E. 1986. A note on Bayes factors for log-linear contingency table models with vague prior information. *J. R. Stat. Soc. B* **48**: 249–250.
- Rao, C. R. 1957. Maximum likelihood estimation for the multinomial distribution. *Sankhyā* **18**: 139–148.
- Rao, C. R. 1961. A study of large sample test criteria through properties of efficient estimates. Part I: Tests for goodness of fit and contingency tables. *Sankhyā Ser. A* **23**: 25–40.
- Rao, C. R. 1963. Criteria of estimation in large samples. *Sankhyā* **25**: 189–206.
- Rao, C. R. 1973. *Linear Statistical Inference and Its Applications*, 2nd ed. Hoboken, NJ: Wiley.
- Rao, C. R. 1982. Diversity: Its measurement, decomposition, apportionment, and analysis. *Sankhyā Ser. A* **44**: 1–22.
- Rao, J. N. K., and A. J. Scott. 1981. The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *J. Am. Stat. Assoc.* **76**: 221–230.
- Rao, J. N. K., and A. J. Scott. 1987. On simple adjustments to chi-square tests with sample survey data. *Ann. Stat.* **15**: 385–397.
- Rao, J. N. K., and D. R. Thomas. 1988. The analysis of cross-classified categorical data from complex sample surveys. *Sociol. Methodol.* **18**: 213–270.
- Rapallo, F. 2003. Algebraic Markov bases and MCMC for two-way contingency tables. *Scand. J. Stat.* **30**: 385–397.
- Rasch, G. 1961. On general laws and the meaning of measurement in psychology. In *Proceedings 4th Berkeley Symposium on Mathematics, Statistics, and Probability*, Vol. 4, ed. J. Neyman. Berkeley,

- CA: University of California Press, pp. 321–333.
- Ravikumar, P., M. Wainwright, and J. Lafferty. 2010. High-dimensional Ising model selection using ℓ_1 -regularized logistic regression *Ann. Stat.* **38**: 1287–1319.
- Read, T. R. C., and N. A. C. Cressie. 1988. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York: Springer-Verlag.
- Reboussin, B. A., and N. S. Ialongo. 2010. Latent transition models with latent class predictors: Attention deficit hyperactivity disorder subtypes and high school marijuana use. *J. R. Stat. Soc. A* **173**: 145–164.
- Rice, N., S. Robone, and P. C. Smith. 2012. Vignettes and health systems responsiveness in cross-country comparative analyses. *J. R. Stat. Soc. A* **175**: 337–369.
- Rice, W. R. 1988. A new probability model for determining exact P -values for 2×2 contingency tables when comparing binomial proportions. *Biometrics* **44**: 1–22.
- Ritov, Y., and Z. Gilula. 1991. The order-restricted RC model for ordered contingency tables: Estimation and testing for fit. *Ann. Stat.* **19**: 2090–2101.
- Robins, J., N. Breslow, and S. Greenland. 1986. Estimators of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* **42**: 311–323.
- Robins, J., A. Rotnitzky, and L. P. Zhao. 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Stat. Assoc.* **90**: 106–121.
- Roeder, K., R. J. Carroll, and B. G. Lindsay. 1996. A semiparametric mixture approach to case-control studies with errors in covariates. *J. Am. Stat. Assoc.* **91**: 722–732.
- Röhmel, J., and U. Mansmann. 1999. Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biomet. J.* **41**: 149–170.
- Rosenbaum, P. R., and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**: 41–55.
- Rosner, B. 1984. Multivariate methods in ophthalmology with application to other paired-data situations. *Biometrics* **40**: 1025–1035.
- Rosner, B. 1989. Multivariate methods for clustered binary data with more than one level of nesting. *J. Am. Stat. Assoc.* **84**: 373–380.
- Rotnitzky, A., and N. P. Jewell. 1990. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* **77**: 485–497.
- Routledge, R. D. 1992. Resolving the conflict over Fisher’s exact test. *Can. J. Stat.* **20**: 201–209.
- Routledge, R. D. 1994. Practicing safe statistics with the mid- P^* . *Can. J. Stat.* **22**: 103–110.
- Roy, S. N., and M. A. Kastenbaum. 1956. On the hypothesis of no “interaction” in a multiway contingency table. *Ann. Math. Stat.* **27**: 749–757.
- Roy, S. N., and S. K. Mitra. 1956. An introduction to some nonparametric generalizations of analysis of variance and multivariate analysis. *Biometrika* **43**: 361–376.
- Royle, J. A., R. M. Dorazio, and W. A. Link. 2007. Analysis of multinomial models with unknown index using data augmentation. *J. Comput. Graph. Stat.* **16**: 67–85.
- Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**: 688–701.
- Rubin, D. B. 1996. Multiple imputation after 18+ years. *J. Am. Stat. Assoc.* **91**: 473–489.
- Rücker, G., G. Schwarzer, J. Carpenter, and I. Olkin. 2008. Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Stat. Med.* **28**: 721–738.
- Rudas, T., C. C. Clogg, and B. G. Lindsay. 1994. A new index of fit based on mixture methods for the analysis of contingency tables. *J. R. Stat. Soc. B* **56**: 623–639.

- Ryan, L. 1992. Quantitative risk assessment for developmental toxicity. *Biometrics* **48**: 163–174.
- Ryan, L. 1995. Comment on article by Liang and Zeger. *Stat. Sci.* **10**: 189–193.
- Ryu, E., and A. Agresti. 2008. Modeling and inference for an ordinal effect size measure. *Stat. Med.* **27**: 1703–1717.
- Samuels, M. L. 1993. Simpson's paradox and related phenomena. *J. Am. Stat. Assoc.* **88**: 81–88.
- Santner, T. J., and D. E. Duffy. 1986. A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **73**: 755–758.
- Santner, T. J., V. Pradhan, P. Senchaudhuri, C. R. Mehta, and A. Tamhane. 2007. Small-sample comparisons of confidence intervals for the difference to two independent binomial proportions. *Comput. Stat. Data An.* **51**: 5791–5799.
- Sato, T. 1989. On the variance estimator for the Mantel–Haenszel risk difference. *Biomet.* **45**: 1323–1324.
- Schaarschmidt, F., M. Sill, and L. A. Hothorn. 2008. Approximate simultaneous confidence intervals for multiple contrasts of binomial proportions. *Biometric J.* **50**: 782–792.
- Schader, M., and F. Schmid. 1990. Charting small sample performance of asymptotic confidence intervals for the binomial parameter p . *Stat. Pap.* **31**: 251–264.
- Schemper, M. 2003. Predictive accuracy and explained variation. *Stat. Med.* **22**: 2299–2308.
- Schluchter, M. D., and K. L. Jackson. 1989. Log-linear analysis of censored survival data with partially observed covariates. *J. Am. Stat. Assoc.* **84**: 42–52.
- Schoenfeld, D., and M. Borenstein. 2005. Calculating the power or sample size for the logistic and proportional hazards models. *J. Stat. Comput. Simul.* **75**: 771–785.
- Scott, A., and C. Wild. 2001. Case–control studies with complex sampling. *Appl. Stat.* **50**: 389–401.
- Seaman, S. R., and S. Richardson. 2004. Equivalence of prospective and retrospective models in the Bayesian analysis of case–control studies. *Biometrika* **91**: 15–25.
- Seeber, G. 2005. Poisson regression. In *Encyclopedia of Biostatistics*, 2nd ed. Chichester, UK: Wiley, pp. 4115–4124.
- Sekar, C. C., and W. E. Derning. 1949. On a method of estimating birth and death rates and the extent of registration. *J. Am. Stat. Assoc.* **44**: 101–115.
- Self, S. G., and K.-Y. Liang. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.* **82**: 605–610.
- Seneta, E., and M. C. Phipps. 2001. On the comparison of two observed frequencies. *Biomet. J.* **43**: 23–43.
- Seneta, E., G. Berry, and P. Macaskill. 1999. Adjustment to Lancaster's mid-P. *Method. Comput. Appl. Prob.* **1**: 229–240.
- Sha, N., M. Vannucci, M. G. Tadesse, P. J. Brown, and I. Dragoni. 2004. Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics* **60**: 812–819.
- Shapiro, S. H. 1982. Collapsing contingency tables: A geometric approach. *Am. Stat.* **36**: 43–46.
- Shuster, J. J. 2010. Empirical vs natural weighted in random effects meta-analysis. *Stat. Med.* **29**: 1259–1265.
- Silvapulle, M. J. 1981. On the existence of maximum likelihood estimators for the binomial response models. *J. R. Stat. Soc. B* **43**: 310–313.
- Simon, G. 1973. Additivity of information in exponential family probability laws. *J. Am. Stat. Assoc.* **68**: 478–482.
- Simon, G. 1974. Alternative analyses for the singly-ordered contingency table. *J. Am. Stat. Assoc.* **69**: 971–976.
- Simonoff, J. 1983. A penalty function approach to smoothing large sparse contingency tables. *Ann.*

- Stat.* **11**: 208–218.
- Simonoff, J. 1986. Jackknifing and bootstrapping goodness-of-fit statistics in sparse multinomials. *J. Am. Stat. Assoc.* **81**: 1005–1111.
- Simonoff, J. S. 1996. *Smoothing Methods in Statistics*. New York: Springer-Verlag.
- Simonoff, J. S. 1998. Three sides of smoothing: Categorical data smoothing, nonparametric regression, and density estimation. *Int. Stat. Rev.* **66**: 137–156.
- Simpson, E. H. 1949. The measurement of diversity. *Nature* **163**: 699.
- Simpson, E. H. 1951. The interpretation of interaction in contingency tables. *J. R. Stat. Soc. B* **13**: 238–241.
- Sison, C. P., and J. Glaz. 1995. Simultaneous confidence intervals and sample size determination for multinomial proportions. *J. Am. Stat. Assoc.* **90**: 366–369.
- Skellam, J. G. 1948. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *J. R. Stat. Soc. B* **10**: 257–261.
- Skene, A. M., and J. C. Wakefield. 1990. Hierarchical models for multicentre binary response studies. *Stat. Med.* **9**: 919–929.
- Skinner, C., and L.-A. Vallet. 2010. Fitting log-linear models to contingency tables from surveys with complex sampling designs: An investigation of the Clogg–Eliason approach. *Sociol. Methods Res.* **39**: 83–108.
- Skrondal, A., and S. Rabe-Hesketh. 2003. Multilevel logistic regression for polytomous data and rankings. *Psychometrika* **68**: 267–287.
- Skrondal, A., and S. Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Slaton, T. L., W. W. Piegorsch, and S. D. Durham. 2000. Estimation and testing with overdispersed proportions using the beta-logistic regression model of Heckman and Willis. *Biometrics* **56**: 125–133.
- Small, K. A. 1987. A discrete choice model for ordered alternatives. *Econometrica* **55**: 409–424.
- Smith, P. W. F., J. J. Forster, and J. W. McDonald. 1996. Monte Carlo exact tests for square contingency tables. *J. R. Stat. Soc. A* **159**: 309–321.
- Smyth, G. K. 2003. Pearson’s goodness of fit statistic as a score test statistic. In *Science and Statistics: A Festschrift for Terry Speed*, ed. D. R. Goldstein, IMS Lecture Notes—Monograph Series, Vol. 40, pp. 115–126, Institute of Mathematical Statistics, Hayward, CA.
- Snell, E. J. 1964. A scaling procedure for ordered categorical data. *Biometrics* **20**: 592–607.
- Sobel, M. E., M. P. Becker, and S. M. Minick. 1998. Origins, destinations, and association in occupational mobility. *Am. J. Sociol.* **104**: 687–721.
- Somers, R. H. 1962. A new asymmetric measure of association for ordinal variables. *Am. Sociol. Rev.* **27**: 799–811.
- Speed, T. 2005. Iterative proportional fitting. In *Encyclopedia of Biostatistics*, 2nd ed. Chichester, UK: Wiley, pp. 2646–2650.
- Spiegelhalter, D. J., and A. F. M. Smith. 1982. Bayes factors for linear and log-linear models with vague prior information. *J. R. Stat. Soc. B* **44**: 377–387.
- Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. B* **64**: 583–639.
- Spitzer, R. L., J. Cohen, J. L. Fleiss, and J. Endicott. 1967. Quantification of agreement in psychiatric diagnosis. *Arch. Gen. Psychiatry* **17**: 83–87.
- Sterne, T. E. 1954. Some remarks on confidence or fiducial limits. *Biometrika* **41**: 275–278.
- Stevens, S. S. 1951. Mathematics, measurement, and psychophysics. In *Handbook of Experimental Psychology*, ed. S. S. Stevens. Hoboken, NJ: Wiley, pp. 1–49.

- Stevens, W. L. 1950. Fiducial limits of the parameter of a discontinuous distribution. *Biometrika* **37**: 117–129.
- Steyerberg, E. W., M. J. C. Eijkemans, F. E. Harrell, and J. D. K. Habbema. 2001. Prognostic modeling with logistic regression analysis. *Med. Decis. Making* **21**: 45–56.
- Stigler, S. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Harvard University Press.
- Stigler, S. 1994. Citation patterns in the journals of statistics and probability. *Stat. Sci.* **9**: 94–108.
- Stigler, S. 1999. *Statistics on the Table*. Cambridge, MA: Harvard University Press.
- Stigler, S. 2002. The missing early history of contingency tables. *Ann. Fac. Sci. Toulouse* **XI**: 563–573.
- Stigler, S. 2008. Karl Pearson’s theoretical errors and the advances they inspired. *Stat. Sci.* **23**: 261–271.
- Stijnen, T., T. Hamza, and P. Özdemir. 2010. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Stat. Med.* **29**: 3046–3067.
- Stiratelli, R., N. Laird, and J. H. Ware. 1984. Random-effects models for serial observations with binary response. *Biometrics* **40**: 1025–1035.
- Stoffer, D. S., D. E. Tyler, and A. J. McDougall. 1993. Spectral analysis for categorical time series: Scaling and the spectral envelope. *Biometrika* **80**: 611–622.
- Stokes, M. E., C. S. Davis, and G. G. Koch. 2012. *Categorical Data Analysis Using SAS*, 3rd ed. Cary, NC: SAS Institute.
- Strawderman, R. L., and M. T. Wells. 1998. Approximately exact inference for the common odds ratio in several 2×2 tables. *J. Am. Stat. Assoc.* **93**: 1294–1307.
- Stuart, A. 1955. A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika* **42**: 412–416.
- Stukel, T. A. 1988. Generalized logistic models. *J. Am. Stat. Assoc.* **83**: 426–431.
- Suissa, S., and J. J. Shuster. 1984. Are uniformly most powerful unbiased tests really best? *Am. Stat.* **38**: 204–206.
- Suissa, S., and J. J. Shuster. 1985. Exact unconditional sample sizes for the 2 by 2 binomial trial. *J. R. Stat. Soc. A* **148**: 317–327.
- Suissa, S., and J. J. Shuster. 1991. The 2×2 matched-pairs trial: Exact unconditional design and analysis. *Biometrics* **47**: 361–372.
- Sundberg, R. 1975. Some results about decomposable (or Markov-type) models for multidimensional contingency tables: Distribution of marginals and partitioning of tests. *Scand. J. Stat.* **2**: 71–79.
- Sutradhar, B. C. 2003. An overview on regression models for discrete longitudinal responses. *Stat. Sci.* **18**: 377–393.
- Swihart, B. J., B. Caffo, and C. Crainiceanu. 2012. A unified approach to modeling multivariate binary data using copulas over partitions. Submitted for publication.
- Tallis, G. 1962. The maximum likelihood estimation of correlation from contingency tables. *Biometrics* **18**: 342–353.
- Tang M. L., N. S. Tang, and I. S. Chan. 2005. Confidence interval construction for proportion difference in small-sample paired studies. *Stat. Med.* **24**: 3565–3579.
- Tango, T. 1998. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Stat. Med.* **17**: 891–908.
- Tanner, M. A., and M. A. Young. 1985. Modelling agreement among raters. *J. Am. Stat. Assoc.* **80**: 175–180.

- Tarone, R. E. 1985. On heterogeneity tests based on efficient scores. *Biometrika* **72**: 91–95.
- Tarone, R. E. 1990. A modified Bonferroni method for discrete data. *Biometrics* **46**: 515–522.
- Tarone, R. E., and J. J. Gart. 1980. On the robustness of combined tests for trends in proportions. *J. Am. Stat. Assoc.* **75**: 110–116.
- Tarone, R. E., J. J. Gart, and W. W. Hauck. 1983. On the asymptotic relative efficiency of certain noniterative estimators of a common relative risk or odds ratio. *Biometrika* **70**: 519–522.
- Tavaré, S., and P. M. E. Altham. 1983. Serial dependence of observations leading to contingency tables, and corrections to chi-squared statistics. *Biometrika* **70**: 139–144.
- Ten Have, T. R. 1996. A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses. *Biometrics* **52**: 473–491.
- Ten Have, T. R., and A. R. Localio. 1999. Empirical Bayes estimation of random effects parameters in mixed effects logistic regression models. *Biometrics* **55**: 1022–1029.
- Ten Have, T. R., and A. Morabia. 1999. Mixed effects models with bivariate and univariate association parameters for longitudinal bivariate binary response data. *Biometrics* **55**: 85–93.
- Theil, H. 1969. A multinomial extension of the linear logit model. *Int. Econ. Rev.* **10**: 251–259.
- Theil, H. 1970. On the estimation of relationships involving qualitative variables. *Am. J. Sociol.* **76**: 103–154.
- Thompson, R., and R. J. Baker. 1981. Composite link functions in generalized linear models. *Appl. Stat.* **30**: 125–131.
- Thompson, W. A. 1977. On the treatment of grouped observations in life studies. *Biometrics* **33**: 463–470.
- Thurstone, L. L. 1927. The method of paired comparisons for social values. *J. Abnormal Social Psychol.* **21**: 384–400.
- Tian, L., T. Cai, M. A. Pfeffer, N. Piankov, P.-Y. Cremieux, and L. J. Wei. 2009. Exact and efficient inference procedure for meta-analysis and its application to the analysis of independent 2×2 tables with all available data but without artificial continuity correction. *Biostatistics* **10**: 275–281.
- Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu. 2003. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Stat. Sci.* **18**: 104–117.
- Tjur, T. 1982. A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scand. J. Stat.* **9**: 23–30.
- Tocher, K. D. 1950. Extension of the Neyman–Pearson theory of tests to discontinuous variates. *Biometrika* **37**: 130–144.
- Toledano, A., and C. Gatsonis. 1996. Ordinal regression methodology for ROC curves derived from correlated data. *Stat. Med.* **15**: 1807–1826.
- Touloumis, A. 2011. Generalized estimating equations for multinomial responses. Ph.D. thesis, University of Florida.
- Train, K. 2009. *Discrete Choice Methods with Simulation*. Cambridge, UK: Cambridge University Press.
- Troendle, J. F., and J. Frank. 2001. Unbiased confidence intervals for the odds ratio of two independent binomial samples with application to case–control data. *Biometrics* **57**: 484–489.
- Tsiatis, A. A. 1980. A note on the goodness-of-fit test for the logistic regression model. *Biometrika* **67**: 250–251.
- Tsodikov, A., and S. Chefo. 2008. Generalized self-consistency: Multinomial logit model and Poisson likelihood. *J. Stat. Plan. Infer.* **138**: 2380–2397.
- Tsutakawa, R. K., and H. Y. Lin. 1986. Bayesian estimation of item response curves. *Psychometrika* **51**: 251–267.
- Turner, H. L., and D. Firth. 2007. gnm: A package for generalized nonlinear models. *R News*. **7**: 8–

- Tutz, G. 1989. Compound regression models for ordered categorical data. *Biomet. J.* **31**: 259–272.
- Tutz, G. 1991. Sequential models in categorical regression. *Comput. Stat. Data An.* **11**: 275–295.
- Tutz, G. 2011. *Structured Regression for Categorical Data*. Cambridge, UK: Cambridge University Press.
- Tutz, G., and W. Hennevogl. 1996. Random effects in ordinal regression models. *Comput. Stat. Data An.* **22**: 537–557.
- Tutz, G., and G. Schauberger. 2012. Visualization of categorical response models from data glyphs to parameter glyphs. Technical report 117, Dept. of Statistics, University of Munich.
- Uebersax, J. S. 1993. Statistical modeling of expert ratings on medical treatment appropriateness. *J. Am. Stat. Assoc.* **88**: 421–427.
- Uebersax, J. S., and W. M. Grove. 1990. Latent class analysis of diagnostic agreement. *Stat. Med.* **9**: 559–572.
- Uebersax, J. S., and W. M. Grove. 1993. A latent trait finite mixture model for the analysis of rating agreement. *Biometrics* **49**: 823–835.
- Umbach, D. M., and C. R. Weinberg. 1997. Designing and analysing case–control studies to exploit independence of genotype and exposure. *Stat. Med.* **16**: 1731–1743.
- Vaeth, M., and E. Skovlund. 2004. A simple approach to power and sample size calculations in logistic regression and Cox regression models. *Stat. Med.* **23**: 1781–1792.
- van der Heijden, P. G. M., and J. de Leeuw. 1985. Correspondence analysis: A complement to log-linear analysis. *Psychometrika* **50**: 429–447.
- van der Heijden, P. G. M., A. de Falguerolles, and J. de Leeuw. 1989. A combined approach to contingency table analysis using correspondence analysis and log-linear analysis. *Appl. Stat.* **38**: 249–292.
- Varin, C., and P. Vidoni. 2006. Pairwise likelihood inference for ordinal categorical time series. *Comput. Stat. Data An.* **51**: 2365–2373.
- Varin, C., N. Reid, and D. Firth. 2011. An overview of composite likelihood methods. *Stat. Sinica* **21**: 5–42.
- Vermunt, J. K. 2003. Multilevel latent class models. *Sociol. Methodol.* **33**: 213–239.
- Vos, P. W., and S. Hudson. 2008. Problems with binomial two-sided tests and the associated confidence intervals. *Austral. New Zealand J. Stat.* **50**: 81–89.
- Wakefield, J. 2004. Ecological inference for 2×2 tables. *J. R. Stat. Soc. A* **167**: 385–445.
- Wald, A. 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Am. Math. Soc.* **54**: 426–482.
- Walker, S. H., and D. B. Duncan. 1967. Estimation of the probability of an event as a function of several independent variables. *Biometrika* **54**: 167–179.
- Walley, P. 1996. Inferences from multinomial data: Learning about a bag of marbles. *J. R. Stat. Soc. B* **58**: 3–34.
- Wang, Z., and T. A. Louis. 2003. Matching conditional and marginal shapes in binary mixed-effects models using a bridge distribution function. *Biometrika* **90**: 765–775.
- Ware, J. H., S. Lipsitz, and F. E. Speizer. 1988. Issues in the analysis of repeated categorical outcomes. *Stat. Med.* **7**: 95–107.
- Warn, D. E., S. G. Thompson, and D. G. Spiegelhalter. 2002. Bayesian random effects meta-analysis of trials with binary outcomes: Methods for the absolute risk difference and relative risk scales. *Stat. Med.* **21**: 1601–1623.
- Warner, S. L. 1963. Multivariate regression of dummy variates under normality assumptions. *J. Am. Stat. Assoc.* **58**: 1054–1063.

- Watson, G. S. 1956. Missing and “mixed up” frequencies in contingency tables. *Biometrics* **12**: 47–50.
- Watson, G. S. 1959. Some recent results in chi-square goodness-of-fit tests. *Biometrics* **15**: 440–468.
- Webb, E. L., and J. J. Forster. 2008. Bayesian model determination for multivariate ordinal and binary data. *Comput. Stat. Data An.* **52**: 2632–2649.
- Wedderburn, R. W. M. 1974. Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61**: 439–447.
- Wedderburn, R. W. M. 1976. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63**: 27–32.
- Wells, M. T. 2010. Optimality results for mid P -values. In *Borrowing Strength: Theory Powering Applications—A Festschrift for Lawrence D. Brown*, eds. J. O. Berger, T. T. Cai, and I. M. Johnstone, IMS Collections, pp. 184–198.
- Wermuth, N. 1976. Model search among multiplicative models. *Biometrics* **32**: 253–263.
- Wermuth, N. 1987. Parametric collapsibility and the lack of moderating effects in contingency tables with a dichotomous response variable. *J. R. Stat. Soc. B* **49**: 353–364.
- Wermuth, N., and S. L. Lauritzen. 1983. Graphical and recursive models for contingency tables. *Biometrika* **70**: 537–552.
- Westfall, P. H., and R. D. Wolfinger. 1997. Multiple tests with discrete distributions. *Am. Stat.* **51**: 3–8.
- Westfall, P. H., and S. S. Young. 1993. *Resampling-Based Multiple Testing: Examples and Methods for p -Value Adjustment*. Hoboken, NJ: Wiley.
- Westfall, P. H., J. F. Troendle, and G. Pennello. 2010. Multiple McNemar tests. *Biometrics* **66**: 1185–1191.
- White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* **50**: 1–26.
- White, A. A., J. R. Landis, and M. M. Cooper. 1982. A note on the equivalence of several marginal homogeneity test criteria for categorical data. *Int. Stat. Rev.* **50**: 27–34.
- Whitehead, J. 1993. Sample size calculations for ordered categorical data. *Stat. Med.* **12**: 2257–2271.
- Whittaker, J. 1990. *Graphical Models in Applied Multivariate Statistics*. Hoboken, NJ: Wiley.
- Whittemore, A. S. 1978. Collapsibility of multidimensional tables. *J. R. Stat. Soc. B* **40**: 328–340.
- Whittemore, A. S. 1981. Sample size for logistic regression with small response probability. *J. Am. Stat. Assoc.* **76**: 27–32.
- Wilks, S. S. 1935. The likelihood test of independence in contingency tables. *Ann. Math. Stat.* **6**: 190–196.
- Wilks, S. S. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **9**: 60–62.
- Williams, D. A. 1975. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31**: 949–952.
- Williams, D. A. 1982. Extra-binomial variation in logistic linear models. *Appl. Stat.* **31**: 144–148.
- Williams, D. A. 1987. Generalized linear model diagnostics using the deviance and single-case deletions. *Appl. Stat.* **36**: 181–191.
- Williams, D. A. 1988. Comments on “The impact of litter effects on dose–response modeling in teratology.” *Biometrics* **44**: 305–308.
- Williams, E. J. 1952. Use of scores for the analysis of association in contingency tables. *Biometrika* **39**: 274–289.
- Williams, O. D., and J. E. Grizzle. 1972. Analysis for contingency tables having ordered response

- categories. *J. Am. Stat. Assoc.* **67**: 55–63.
- Williams, P. L. 2005. Trend tests for counts and proportions. In *Encyclopedia of Biostatistics*, 2nd ed. Hoboken, NJ: Wiley, pp. 5516–5527.
- Wilson, E. B. 1927. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.* **22**: 209–212.
- Wilson, E. B., and J. Worcester. 1943. The determination of L.D. 50 and its sampling error in bioassay. *Proc. Natl. Acad. Sci. USA* **29**: 79–85.
- Witten, D. M., and R. Tibshirani. 2011. Penalized classification using Fisher's linear discriminant. *J. R. Stat. Soc. B* **73**: 753–772.
- Wong, G. Y., and W. M. Mason. 1985. The hierarchical logistic regression model for multilevel analysis. *J. Am. Stat. Assoc.* **80**: 513–524.
- Wong, R. S.-K. 2010. *Association Models*. Thousand Oaks, CA: Sage, Publications.
- Woolf, B. 1955. On estimating the relation between blood group and disease. *Ann. Human Genet. (London)* **19**: 251–253.
- Woolson, R. F., and W. R. Clarke. 1984. Analysis of categorical incomplete longitudinal data. *J. R. Stat. Soc. A* **147**: 87–99.
- Wu, M. C., L. Zhang, Z. Wang, D. C. Christiani, and X. Lin. 2009. Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics* **25**: 1145–1151.
- Xie, Y. 1992. The log-multiplicative layer effect model for comparing mobility tables. *Am. Sociol. Rev.* **57**: 380–395.
- Yang, I., and M. P. Becker. 1997. Latent variable modeling of diagnostic accuracy. *Biometrics* **53**: 948–958.
- Yang, M., H. Goldstein, and A. Heath. 2000. Multilevel models for repeated binary outcome: Attitudes and voting over the electoral cycle. *J. R. Stat. Soc. A* **163**: 49–62.
- Yang, M.-C., D.-W. Lee, and J. T. G. Hwang. 2004. Equivalence of the mid p-value and the expected p-value for testing equality of two balanced binomial proportions. *J. Stat. Plan. Infer.* **126**: 273–280.
- Yates, F. 1934. Contingency tables involving small numbers and the χ^2 test. *J. R. Stat. Soc. Suppl.* **1**: 217–235.
- Yates, F. 1948. The analysis of contingency tables with grouping based on quantitative characters. *Biometrika* **35**: 176–181.
- Yates, F. 1955. The use of transformations and maximum likelihood in the analysis of quantal experiments involving two treatments. *Biometrika* **42**: 382–403.
- Yates, F. 1984. Tests of significance for 2×2 contingency tables. *J. R. Stat. Soc. A* **147**: 426–463.
- Yee, T. W., and C. J. Wild. 1996. Vector generalized additive models. *J. R. Stat. Soc. B* **58**: 481–493.
- Yerushalmy, J. 1947. Statistical problems in assessing methods of medical diagnosis, with special reference to x-ray techniques. *Public Health Rep.* **62**: 1432–1449.
- Yule, G. U. 1900. On the association of attributes in statistics. *Philos. Trans. R. Soc. London Ser. A* **194**: 257–319.
- Yule, G. U. 1903. Notes on the theory of association of attributes in statistics. *Biometrika* **2**: 121–134.
- Yule, G. U. 1906. On a property which holds good for all groupings of a normal distribution of frequency for two variables, with application to the study of contingency tables for the inheritance of unmeasured qualities. *Proc. Roy. Soc. A* **77**: 324–336.
- Yule, G. U. 1912. On the methods of measuring association between two attributes. *J. R. Stat. Soc.*

75: 579–642.

- Zeger, S. L., and M. R. Karim. 1991. Generalized linear models with random effects: A Gibbs sampling approach. *J. Am. Stat. Assoc.* **86**: 79–86.
- Zeger, S. L., and Qaqish, B. 1988. Markov regression models for time series: A quasi-likelihood approach. *Biometrics* **44**: 1019–1031.
- Zeger, S. L., K.-Y. Liang, and P. S. Albert. 1988. Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**: 1049–1060.
- Zelen, M. 1971. The analysis of several 2×2 contingency tables. *Biometrika* **58**: 129–137.
- Zelen, M. 1991. Multinomial response models. *Comput. Stat. Data An.* **12**: 249–254.
- Zellner, A., and P. E. Rossi. 1984. Bayesian analysis of dichotomous quantal response models. *J. Econometrics* **25**: 365–393.
- Zelterman, D. 1987. Goodness-of-fit tests for large sparse multinomial distributions. *J. Am. Stat. Soc.* **82**: 624–629.
- Zermelo, E. 1929. Die Berechnung der Turnier-Ergebnisse als ein Maximum problem der Wahrscheinlichkeitsrechnung. *Math. Z.* **29**: 436–460.
- Zhang, H. 1998. Classification trees for multiple binary responses. *J. Am. Stat. Soc.* **93**: 180–193.
- Zhang, H., and B. H. Singer. 2010. *Recursive Partitioning and Applications*, 2nd ed. New York: Springer.
- Zhang, H. H., J. Ahn, X. Lin, and C. Park. 2006. Gene selection using support vector machines with non-convex penalty. *Bioinformatics* **22**: 88–95.
- Zhang, Y., and J. S. Liu. 2007. Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics* **39**: 1167–1173.
- Zhang, Y., B. Jiang, J. Zhu, and J. S. Liu. 2011. Bayesian models for detecting epistatic interactions from genetic data. *Ann. Human Genetics* **75**: 183–193.
- Zhao, L. P., and R. L. Prentice. Correlated binary regression using a quadratic exponential model. *Biometrika* **77**: 642–648.
- Zheng, B., and A. Agresti. 2000. Summarizing the predictive power of a generalized linear model. *Stat. Med.* **19**: 1771–1781.
- Zhu, J., and T. Hastie. 2004. Classification of gene microarrays by penalized logistic regression. *Biostatistics* **46**: 505–510.
- Zhu, Y., and N. Reid. 1994. Information, ancillarity, and sufficiency in the presence of nuisance parameters. *Can. J. Stat.* **22**: 111–123.
- Zipunnikov, V., and J. Booth. 2012. Closed form GLM cumulants and GLMM fitting with a SQUAR-EM-LA₂ algorithm. Submitted for publication.
- Zocchi, S. S., and A. C. Atkinson. 1999. Optimum experimental designs for multinomial logistic models. *Biometrics* **55**: 437–444.

Author Index

- Achten, J.
Agresti, A.
Ahn, J.
Aitchison, J.
Aitken, C.
Aitkin, M.
Albert, A.
Albert, J.
Albert, P.
Allison, P.
Altham, P.
Amara, I.
Amemiya, T.
An, M.-W.
Andersen, A.
Andersen, E.
Anderson, C.
Anderson, D.
Anderson, J.
Anderson, R.
Anderson, T.
Aranda-Ordaz, F.
Armitage, P.
Ashford, J.
Asmussen, S.
Aston, J.
Atkinson, A.
Azzalini, A.

Böckenholt
Bühlmann, P.
Baglivo, J.
Baker, F.
Baker, R.
Baker, S.
Banerjee, C.
Banerjee, S.
Baptista, J.
Barnard, G.
Bartholomew, D.
Bartlett, M.
Bartolucci, F.
Becker, M.
Beder, J.
Bedrick, E.
Begg, C.
Beggs, S.

Beitler, P.
Benedetti, J.
Benichou, J.
Benamini, Y.
Bennett, J.
Benzéri, J.-P.
Berger, R.
Bergsma, W.
Berkson, J.
Berry, G.
Berry, S.
Bertaccini, B.
Besag, J.
Bhadra, D.
Bhapkar, V.
Bickel, P.
Billingsley, P.
Bini, M.
Birch, M.
Bishop, Y.
Blaker, H.
Blanchard, G.
Bliss, C.
Bloch, D.
Blyth, C.
Bock, R. D.
Bonney, G.
Boos, D.
Booth, J.
Borenstein, M.
Boschloo, R.
Bowden, D.
Bowker, A.
Bowman, A.
Box, J. F.
Bradford Hill, A.
Bradley, R.
Branco, J.
Brandt, A. E.
Branscum, A.
Brazzale, A.
Breiman, L.
Breslow, N.
Brier, S.
Brockmann, J.
Brooks, S.
Brass, I.
Brown, C.

Brown, L.
Brown, M.
Brown, P.
Browne, W.
Brusco, M.
Budtz-Jørgensen
Bull, S.
Buonaccorsi, J.
Bumham, K.
Burr, D.
Burridge, J.
Butler, R.
Byar, D.

Caffo, B.
Cai, T.
Cameron, A.
Campbell, I.
Capanu, M.
Capozzoli, M.
Cardell, S.
Carey, V.
Carlin, B.
Carlin, J.
Carpenter, J.
Carroll, R.
Casella, G.
Catalano, P.
Caussinus, H.
Cerioli, A.
Chafaï, D.
Chaganty, N.
Chaloner, K.
Chamberlain, G.
Chambers, E.
Chambers, R.
Chan, I.
Chao, A.
Chapman, D.
Chapman, P.
Chatterjee, N.
Chefo, S.
Chen, M.-H.
Chen, Z.
Cheng, P.
Chib, S.
Chinchilli, V.
Choi, J.
Choulakian, V.

Christensen, R.
Chuang, C.
Chuang-Stein, C.
Clarke, W.
Clayton, D.
Clogg, C.
Clopper, C.
Cochran, W.
Cohen, A.
Cohen, J.
Coleman, J.
Collett, D.
Collins, L.
Colombi, R.
Conaway, M.
Congdon, P.
Connell, D.
Consonni, G.
Conti, D.
Cook, D.
Cooper, M.
Copas, J.
Corcoran, C.
Cordeiro, G.
Cormack, R.
Cornfield, J.
Costa, M.
Coull, B. A.
Cox, C.
Cox, D. R.
Crainiceanu, C.
Cramér, H.
Cramer, J.
Cremieux, P.-Y.
Cressie, N.
Crook, J.
Croon, M.
Crouch, B.
Crowder, M.
Dahinden, C.
Dahm, P.
Daniels, M.
Darnell, R.
Darroch, J.
Das Gupta, S.
DasGupta, A.
David, H. A.
Davis, C.

Davis, G.
Davis, L.
Davison, A.
Dawid, A. P.
Dawson, R.
Day, N.
de Falguerolles, A.
de Leeuw, J.
De Menezes, R.
de Rooij, M.
Dellaportas, P.
Demétrio, C.
Deming, W. E.
Dersimonian, R.
Deville, J.-C.
Dey, D.
Diaconis, P.
Dickey, J.
Diggle, P.
Dillon, W.
Dittrich, R.
Dobra, A.
Doksum, K.
Doll, R.
Dong, J.
Donner, A.
Doolittle, M.
Dorazio, R.
Doss, H.
Dragoni, I.
Drost, F.
Drton, M.
Ducharme, G.
Dudbridge, F.
Dudoit, S.
Duffy, D.
Duncan, D.
Duncan, O. D.
Dunson, D.
Dupont, W.
Durham, S.
Dyke, G.
Edwardes, M.
Edwards, A.
Edwards, D.
Efron, B.
Eguchi, S.
Eijkemans, M.

Eisenmann, S.

Ekholm, A.

Emerson, J.

Endahl, L.

Endicott, J.

Eriksson, N.

Erosheva, E.

Escoufier, Y.

Espeland, M.

Everitt, B.

Ewings, P.

Fagerland, M.

Fahrmeir, L.

Fan, J.

Fan, X.

Fan, Y.

Farcomeni, A.

Farewell, V.

Farrington, C.

Fay, M.

Fay, R.

Fechner, G.

Fieger, A.

Fienberg, S.

Finney, D.

Firth, D.

Fischer, G.

Fisher, M.

Fisher, R. A.,

Fitzmaurice, G.

Fitzpatrick, S.

Fleiss, J.

Fokianos, K.

Follman, D.

Forcina, A.

Forster, J.

Fowlkes, E.

Fraley, C.

Francis, B.

Francom, S.

Frank, J.

Franklin, J.

Fraser, D.

Freeman, D.

Freeman, G.

Freeman, J.

Freeny, A.

Freidlin, B.

Fridlyand, J.

Friedl, H.

Friedman, J.

Friendly, M.

Frome, E.

Funo, E.

Gabriel, R.

Gail, M.

Gange, S.

Gart, J.

Gaskins, R.

Gasko, M.

Gastwirth, J.

Gatsonis, C.

Gelfand, A.

Gellad, W.

Gelman, A.

Geng, Z.

Genkin, A.

Genter, F.

Geyer, C.

Ghosh, A.

Ghosh, B. K.

Ghosh, D.

Ghosh, M.

Ghosh, S.

Gibbons, R.

Gilbert, G.

Gilbert, P.

Gillings, D.

Gilmour, A.

Gilula, Z.

Gini, C.

Giordano, S.

Glaz, J.

Gleser, L.

Gloeckler, L.

Glonek, G.

Godambe, V.

Goetghebeur, E.

Gokhale, D.

Goldberg, R.

Goldstein, H.

Good, I. J.

Goodman, L. A.

Gottard, A.

Gourieroux, C.

Gover, T.

Graubard, B.
Gray, R.
Green, P.
Greenacre, M.
Greenberg, E.
Greene, G.
Greene, W.
Greenland, S.
Greenwood, M.
Greenwood, P.
Grego, J.
Grevstad, N.
Griswold, M.
Grizzle, J.
Gross, S.
Grove, W.
Gueorguieva, R.
Guerrero, V.
Gunel, E.
Guo, G.
Gurland, J.

Härdle, W.
Haas, J.
Habbema, J.
Haber, M.
Haberman, S.
Haenszel, W.
Hagenaars, J.
Hald, A.
Haldane, J. B. S.
Hall, P.
Halperin, M.
Halton, J.
Halvorsen, K.
Hamada, M.
Hamdan, M.
Hamza, T.
Hand, D.
Handelman, S.
Hanfelt, J.
Hanley, J.
Hansen, L.
Hanson, T.
Harrell, F.
Harrington, D.
Hartley, H. O.
Hartzel, J.
Haseman, J.

Hashemi, L.
Haslen, S.
Hastie, T.
Hatzinger, R.
Hauck, W.
Hausman, J.
Hawkins, D.
Heagerty, P.
Heath, A.
Heckman, J.
Hedeker, D.
Heim, R.
Heinen, T.
Heinze, G.
Heiser, W.
Held, L.
Henley, W.
Hennevogl, W.
Henretta, J.
Hensher, D.
Herring, A.
Heumann, C.
Heyde, C.
Heyman, E.
Hilbe, J.
Hill, J.
Hinde, J.
Hinkley, D.
Hirji, K.
Hirotsu, C.
Hirschfeld, H. O.
Hitchcock, D.
Hoaglin, D.
Hobert, J.
Hochberg, Y.
Hodges, J.
Hoem, J.
Hoeting, J.
Hoff, P.
Holford, T.
Holland, P.
Hollander, M.
Holmes, C.
Holt, A.
Holtbrugge, W.
Hook, E.
Hosmer, D.
Hosmer, T.

Hothom, L. A.

Hout, M.

Howard, J.

Hsieh, F.

Hsu, J.

Hu, B.

Hubert, L.

Hudson, S.

Hui, S.

Hunt, L.

Hwang, G.

Ialongo, N.

Ibrahim, J.

Imai, K.

Imrey, P.

Ireland, C.

Irwin, J.

Jackson, K.

Jensen, S.

Jewell, N.

Joe, H.

Johnson, B.

Johnson, M.

Johnson, N.

Johnson, R.

Johnson, V.

Johnson, W.

Jones, B.

Jones, K.

Jones, L.

Jones, M.

Jorgensen, M.

Joutard, C.

Junker, B.

Jørgensen, B.

Kalbfleisch, J.

Kalisch, M.

Kallenberg, C.

Karim, M.

Kass, G.

Kastenbaum, M.

Kastner, C.

Kateri, M.

Katzenbeisser, W.

Kauermann, G.

Kaufman, L.

Kaufmann, H.

Kawaguchi, A.
Kedem, B.
Kelderman, H.
Kemp, A.
Kempthorne, O.
Kendall, M. G.
Kenward, M.
Kezouh, A.
Khamis, H.
Khuri, A.
Kiefer, N.
Kim, D.
Kim, K.
Kim, S.
King, G.
King, T.
Klingenberg, B.
Kneib, T.
Knott, M.
Knuiman, M.
Koch, G. G.
Koehler, K.
Koopman, P.
Korn, E.
Kosmidis, I.
Kotz, S.
Kraemer, H.
Krampe, A.
Kreiner, S.
Kruskal, W.
Ku, H.
Kuha, J.
Kuhnt, S.
Kulinskaya, E.
Kullback, S.
Kuo, L.
Kupper, L.
Kupperman, M.
Kuritz, S.
Läärä, E.
La Rocca, L.
Laake, P.
Lachin, J.
Lafferty, J.
Laird, N.
Lambert, D.
Lambert, P.
Lancaster, H. O.

Landau, S.
Landis, J. R.
Landrum, M.
Landwehr, J.
Lang, J.
Lanza, S.
Laplace, P. S.
Larntz, K.
Larsen, K.
Larsen, M.
Larson, M.
Laud, P.
Lauritzen, S.
LaVange, L.
Lawless, J.
le Cessie, S.
Lee, A.
Lee, D.
Lee, I.
Lee, S.
Lee, Y.
Leecaster, M.
Leese, M.
Lefkopoulou, M.
Lehmann, E.
Lehnen, R.
Lele, S.
Lemeshow, S.
Lemke, J.
Leonard, T.
Lepage, Y.
Lesaffre, E.
Lesperance, M.
Levin, B.
Levina, E.
Lewis, T.
Li, D.
Li, P.
Li, X.
Liang, K.-Y.
Liao, J.
Lin, C.-Y.
Lin, H.
Lin, H.-M.
Lin, X.
Lindley, D.
Lindsay, B.
Lindsey, J.

Lindstrom, M.
Link, W.
Liou, M.
Lipsitz, S.
Little, R.
Liu, D.
Liu, I.-M.
Liu, J.
Liu, Q.
Lloyd, C.
Localio, R.
Loh, W.-Y.
Longford, N.
Loughin, T.
Louis, T.
Lovison, G.
Luce, R.
Lumley, T.
Lupparelli, M.
Lv, J.
Lydersen, S.
Lyles, R.
Müller, P.
Madansky, A.
Maddala, G.
Madigan, D.
Magidson, J.
Mai, Q.
Maiste, P.
Makov, U.
Mallick, B.
Manning, G.
Mansmann, U.
Mantel, N.
Marchetti, G.
Martín Andrés, A.
Mason, W.
Massam, H.
Matthews, J.
McCloud, P.
McCullagh, P.
McCulloch, C.
McCulloch, R.
McCutcheon, A.
McDonald, J.
McDougall, A.
McFadden, D.
McGee, D.

McKelvey, R.
McLachlan, G.
McNeil, B.
McNemar, Q.
McSweeney, L.
Mee, R.
Meeden, G.
Mehrotra, D.
Mehta, C.
Meier, L.
Menard, S.
Meng, R.
Meulman, J.
Michailidis, G.
Miettinen, O.
Miller, J.
Miller, M.
Min, Y.
Minick, S.
Mirkin, B.
Mitchell, G.
Mitra, S.
Mittal, Y.
Molenaar, I.
Molenberghs, G.
Monfort, A.
Moore, C.
Moore, D. F.
Moore, D. S.
Morabia, A.
Moreno, E.
Morgan, B.
Morris, C.
Morris, K.
Mosteller, F.
Moustaki, I.
Muenz, L.
Mukherjee, B.
Mutanen, P.

Nair, V.
Nam, J.
Nandrum, B.
Natarajan, R.
Natarajan, S.
Nelder, J.
Neuhaus, J.
Newcombe, R.
Newton, M.

Neyman, J.
Nicolaou, A.
Nikulin, M.
Noda, A.
Noe, D.
Normand, S.-L.
Norris, J.
Ntzoufras, I.
Nurminen, M.

O'Brien, P.
O'Brien, R.
O'Brien, S.
O'Gorman, T.
O'Hagan, A.
Ochi, Y.
Ohman-Strickland, P.
Olive, D.
Olivier, D.
Olkin, I.
Olshen, R.
Oosterhoff, J.
Osius, G.
Overton, S.
Owen, A.
Ozdemir, P.

Pack, S.
Pagano, M.
Paige, R.
Paik, M.
Palmgren, J.
Palta, M.
Papaioannou, T.
Park, C.
Park, M.
Park, T.
Parsons, N.
Parzen, E.
Parzen, M.
Patefield, W.
Patel, N.
Patnaik, P.
Patterson, H.
Paul, S.
Pavlides, M.
Pawitan, Y.
Pearl, J.
Pearson, E. S.

Pearson, K.
Peduzzi, P.
Pendergast, J.
Pennello, G.
Pepe, M.
Periyakoil, V.
Perlman, M.
Peters, D.
Petersen, J.
Peterson, B.
Peto, R.
Pfeffer, M.
Philips, P.
Phipps, M.
Piankov, N.
Piccarreta, R.
Pickles, A.
Piegorsch, W.
Pierce, D.
Pierce, G.
Pike, M.
Pires, A.
Plackett, R.
Pledger, S.
Poisson, S.
Pollock, K.
Poison, N.
Portier, C.
Potter, D.
Powers, W.
Pratt, J.
Pregibon, D.
Prentice, R.
Presnell, B.
Press, S.
Pyke, R.

Qaqish, B.
Qin, J.
Quetelet, A.
Quine, M.
Quintana, F.

Røhmel, U.
Rücker, G.
Rabbee, N.
Rabe-Hesketh, S.
Racine, A.
Radelet, M.

Rae, A.
Raftery, A.
Rao, C. R.
Rao, J. N. K.
Rapallo, F.
Rasch, G.
Ratcliff, D.
Ravikumar, P.
Read, T.
Reboussin, B.
Regal, R.
Reid, N.
Rice, N.
Rice, W.
Richardson, S.
Richardson, T.
Ridout, M.
Rinaldo, A.
Ritov, Y.
Rizopoulos, D.
Robins, J.
Robone, S.
Roeder, K.
Rohde, C.
Rojek, D.
Romano, J.
Rosenbaum, P.
Rosner, B.
Rossi, P.
Rotnitzky, A.
Rousseeuw, P.
Routledge, R.
Roy, S.
Royle, J.
Rubin, D.
Rubin, H.
Rubinstein, L.
Rudas, T.
Rundell, P.
Ryan, L.
Ryu, E.
Särndal, C.-E.
Sabai, C.
Sackrowitz, H.
Samuels, M.
Sands, B.
Santner, T.
Sato, T.

Saunders, I.
Sautory, O.
Scarpa, B.
Schaarschmidt, F.
Schader, M.
Schafer, D.
Schauberger, G.
Schemper, M.
Scherer, P.
Schluchter, M.
Schmid, F.
Schoenfeld, D.
Schumacher, M.
Schwartz, T.
Schwartzman, A.
Schwarzer, G.
Scott, A.
Seaman, S.
Searle, S.
Seeber, G.
Self, S.
Senchaudhuri, P.
Seneta, E.
Sha, N.
Shao, J.
Shao, Q.-M.
Shapiro, S.
Sharp, R.
Shen, S.
Shih, Y.-S.
Shihadeh, E.
Shoemaker, A. C.
Shuster, J.
Sill, M.
Silva Mato, A.
Silvapulle, M.
Silvey, S.
Simon, G.
Simonoff, J.
Simpson, E. H.
Singer, B.
Sinha, B.
Sinha, D.
Sison, C.
Skellam, J.
Skene, A.
Skinner, C.
Skovlund, E.

Skrondal, A.
Slaton, T.
Small, K.
Smith, A.
Smith, P.
Smyth, G.
Snedecor, G.
Snell, J.
Sobel, M.
Somes, G.
Sowden, R.
Speed, T.
Speizer, F.
Spiegelhalter, D.
Spiessens, B.
Spitzer, R.
Stahl, D.
Stallard, N.
Starmer, F.
Steel, D.
Stephan, F.
Stern, H.
Sterne, T.
Stevens, S. S.
Stevens, W.
Steyerberg, E.
Stigler, S.
Stijnen, T.
Still, H.
Stiratelli, R.
Stoffer, D.
Stokes, M.
Stone, C.
Strawderman, R.
Stuart, A.
Stukel, T.
Sturmfels, B.
Subramanian, S.
Suissa, S.
Sundberg, R.
Sutradhar, B.
Swihart, B.
Tadesse, M.
Tallis, G.
Tang, M.
Tang, N.
Tango, T.
Tanner, M.

Tarone, R.
Tavaré, S.
Taylor, J.
Ten Have, T.
Terry, M.
Theil, H.
Thomas, D.
Thompson, R.
Thompson, S.
Thompson, W.
Thurstone, L.
Tian, L.
Tibshirani, R.
Titterington, D.
Tjur, T.
Tocher, K.
Toledano, A.
Touloumis, A.
Train, K.
Trivedi, P.
Troendle, J.
Trognon, A.
Tsiatis, A.
Tsodikov, A.
Tsonaka, R.
Tsutakawa, R.
Turner, H.
Tutz, G.
Tyler, D.
Uebersax, J.
Umbach, D.
Vaeth, M.
Vallet, L.-A.
van de Geer, S.
van der Heijden, P.
van Dyk, D.
van Houwelingen, J.
Vannucci, M.
Varin, C.
Verbeke, G.
Vermunt, J.
Vidoni, P.
Vieira, A.
Vos, P.
Wainer, H.
Wainwright, M.
Wakefield, J.

Wald, A.
Walker, S.
Walley, P.
Wang, C.
Wang, S.
Wang, X.
Wang, Z.
Ware, J.
Warn, D.
Warner, S.
Wasserman, S.
Watson, G.
Weakliem, D.
Webb, E.
Wedderburn, R.
Wei, L. J.
Weinberg, C.
Weir, B.
Weisberg, S.
Wells, M.
Wermuth, N.
Westcott, M.
Westfall, P.
White, A.
White, H.
Whitehead, J.
Whittaker, J.
Whittemore, A.
Wild, C.
Wilks, S.
Williams, D.
Williams, E.
Williams, O. D.
Williams, P.
Williamson, J.
Wilson, E. B.
Wilson, J.
Wilson, S.
Winner, L.
Wise, D.
Witten, D.
Wolfe, R.
Wolfinger, R.
Wong, G.
Wong, R.
Woolf, B.
Woolson, R.
Worcester, J.

Wu, C. F. J.

Wu, M.-M.

Xie, Y.

Yaari, G.

Yang, I.

Yang, M.

Yang, M.-C.

Yates, F.

Yee, T.

Yekutieli, D.

Yerushalmi, J.

York, J.

Young, M.

Young, S.

Yu, J.-T.

Yuan, Y.

Yule, G. U.

Zavoina, W.

Zeger, S.

Zelen, M.

Zellner, A.

Zelterman, D.

Zeng, L.

Zermelo, E.

Zhang, H.

Zhang, T.

Zhang, Z.

Zhao, H.

Zhao, L.

Zheng, B.

Zhu, J.

Zhu, Y.

Zocchi, S.

Zweifel, J.

Example Index

Abortion altitudes
AIDS and AZT use
AIDS and government measures
Air pollution and respiratory illness
Alcohol, cigarettes, marijuana
Alligator food choice
Appendix pain
Assisted living enrollment
Auto accidents and seat belts

Baseball home team advantage
Baseball results
Basketball
 Kobe Bryant shooting
 Rajon Rondo assists
 Ray Allen shooting
Beetle mortality
Belief in God by educational level
Belief in heaven
Breast cancer and tamoxifen
Buchanan Presidential votes
Butterfly ballot

Cancer and smoking case-control study
Cancer of larynx by treatment
Cancer remission and labeling index
Cannabis use and mother's age
Carcinoma diagnoses by pathologists
Cardiovascular disease and teeth brushing
Carp malformation and lead pollution
Child respiratory illness and maternal smoking
Children's care for mother
Cholesterol and psyllium
Clinical trial for fungal infections
Coffee purchases
Cola taste tests
Condom use
Coronary death rates, smoking, and age
Credit scoring
Crossover drug comparison
Crossover trial for dysmenorrhea

Death penalty
Death penalty and race
Death rates and Simpson's paradox
Developmental toxicity study
Diarrhea and an antibiotic
Divorce grounds

Draft position and all-star
Driving after consuming alcohol
Dumping severity
Dysmenorrhea crossover trial

Educational aspirations and family income
Endometrial cancer grade
Esophageal cancer and alcohol consumption
Evapotranspiration rates

Fish hatching

Gambler's ruin
Gene-environment interactions
Genomics

Gestation length and infant survival
Global warming attitudes

Golf putting
Government spending
Graduate school admissions

Berkeley
Florida

Graham Greene Russian roulette
Gun-related deaths by nation

Halothane study
Happiness and political ideology
Happiness and traumatic events

Heart attacks and aspirin use
Heart catheterization and race

Heart disease
blood pressure
smoking
snoring

Heart valve operation survival
Heaven and hell
Hepatitis capture-recapture
Home ownership
Homicide victim frequency

Homosexual marriage and party ID
Homosexual marriage and religious fundamentalism
Homosexual sex and premarital sex

Horseshoe crab mating

binary modeling
classification tree
count modeling
discriminant analysis
generalized additive model
linear probability model

Infant malformation and alcohol consumption

Insomnia clinical trial

Job satisfaction

- by age
- by income
- by race
- predicting

Journal citations

Kyphosis risk factors

Leading crowd membership/attitudes

Lung cancer and smoking

Lung cancer clinical trial

Lung cancer survival

Malformation and alcohol consumption

Marginal vs. conditional associations

Marital status causal models

Medical diagnoses

Mendel's theories

Mental depression

Mental impairment and parents' SES

Menu pricing

Meta-analysis

Migration

Missing people

Motif discovery

Movie ratings

Multicenter clinical trial

Multiple sclerosis evaluations

Murder and race

Murder rates by gender and race

Myers-Briggs personality scales

Myocardial infarction and diabetes

NBA basketball predicted probabilities

Netflix prize

Obesity by gender and time

Occupational aspirations

Occupational mobility

Oral contraceptive use

Oxford/Cambridge boat race

Pain after surgery

Party ID and gender

Party ID and race

Penicillin in rabbits

Pig farmer survey

Pneumonia infections in calves

Polarized opinions

Political party ID and attitudes

Political party ID, gender and race

Pregnancy rates

Premarital and extramarital sex

Premarital sex and birth control

Presidential voting

2004 and 2008 elections

Buchanan and butterfly ballot

election clustering

election poll

stem cell research

Prime minister evaluation

Prison rates by nation

Promotion discrimination

Prostate cancer diagnostic test

Protozoan and poison dose

Prussian army and mule kicks

Psychiatric diagnosis and prescribed drug

Quality of life

Rap music liking

Regional migration

Satisfaction with housing

Seat belts and injury

Sexual intercourse frequency

Sexual orientation and party ID

Shopping choice

Silicon wafer imperfections

Simpson's paradox

baseball batting averages

death penalty

death rates

graduate school admissions

Siskel and Ebert movie ratings

Skin damage and leprosy

Snoring and heart disease

Snowshoe hare capture–recapture

Sore throat after surgery

Space shuttle

Stem cell research

Taxes for environment

Tea tasting

Tennis results

Teratology overdispersion

Titanic survival and gender

Toenail infection clinical trial

Traffic deaths and seat-belt use

Trauma patient survival

Travel credit card

Urn sampling in clinical trial

Vegetarianism

World Cup odds

Subject Index

Adding constants to cell counts
Adjacent-categories logit model
Adjusted response variable
Agreement
Agresti-Coull confidence interval
AIC
lasso special case
Alternating logistic regressions
Amalgamation paradox
Ancillarity
Arc sine transformation
Association factor
Association measures
Association models
row and column effects
Asymptotics
 delta method
 higher-order
Attributable risk
Average causal effect
Backward elimination
Bagging
Baseline-category logit model
 adjacent category logits
 Bayesian fitting
 conditional independence
 discrete-choice model
 exponential family
 likelihood function
 matched pairs
 matched sets
 random effects
 references
 sufficient statistics
Bayesian inference
 binary regression models
 CDA history
 comparing proportions
 equal-tail interval
 GLMs
 highest posterior density interval
 introduction
 loglinear models
 marginal models
 model averaging
 model checking

- multinomial models
- multivariate responses
- posterior interval
- two-way tables

Bernoulli trials

- correlated

Best asymptotically normal (BAN)

Beta distribution

- prior for binary regression
- prior for binomial parameter
- prior for comparing proportions

Beta-binomial models

- beta-binomial distribution

Between-subject effects

Bias of reasonable discrete tests

Bias reduction

- generalized linear models
- logistic regression

Bias/variance tradeoff

- classification trees
- discriminant analysis
- estimation
- kernel smoothing
- penalized likelihood

BIC

Binomial distribution

- Bayesian inference
- confidence intervals
- exponential family form
- inference
- likelihood function
- likelihood-ratio test
- moment generating function
- properties
- score test
- small-sample inference

Binomial GLMs

- deviance
- likelihood equations

Biserial correlation

Bonferroni multiple comparisons

- binomial parameters
- discrete adjustment
- multinomial parameters
- multiple testing

Bradley-Terry model

- quasi-symmetry model
- sufficient statistics

Brandt–Snedecor formula

Breslow–Day test

Calibration

Canonical correlation model

Canonical link function

Capture–recapture modeling

Case-control study

- logistic regression

- matched pairs

- odds ratio estimation

Cauchit link function

Chi-squared distribution

- moment generating function

- partitioning

- properties

Chi-squared test

- adequacy of approximation

- derivation

- independence

- logistic goodness-of-fit

- loglinear goodness-of-fit

- multinomial goodness-of-fit

- noncentral distribution

- power

- sparse data asymptotics

Classification

- discriminant analysis

- high dimensions

- logistic regression

- multiple categories

- tree-structured

Classification table

Classification tree

- vs. logistic regression

Clopper–Pearson confidence interval

Cluster analysis

Cluster sampling

Clustered data

- contingency tables

Cochran's Q

Cochran–Armitage trend test

Cochran–Mantel–Haenszel test

- generalized

- McNemar test connection

Cochran, William, contributions to CDA

Collapsibility

- difference of proportions

- odds ratio

- relative risk
- Comparing measures
- Comparing models
 - deviances
 - Pearson statistic
 - sparse data
- Complementary log-log model
 - continuation ratios
 - ordinal response
- Complete separation
- Complete symmetry
- Composite likelihood
- Concentration coefficient
- Concordance index
- Concordant pair
- Conditional independence
 - binary response tests
 - graphs
 - loglinear model
 - loglinear model test
 - marginal independence
 - multinomial response tests
 - ordinal loglinear test
 - small-sample tests
 - test power
- Conditional inference
 - contingency tables
 - controversy
 - logistic regression
- Conditional logistic regression
 - binary matched pairs
 - matched case-control
- Conditional logit model, *see* Discrete-choice model
- Conditional ML estimation
 - between-cluster effect
 - odds ratio
 - random effects comparison
- Conditional symmetry
- Confidence intervals
 - inverting tests
 - profile likelihood
 - score-test-based
 - simultaneous
 - small-sample
- Confounding
- Conjugate mixture models
- Conservatism in discrete inference
- Constraints on parameters

Contingency coefficient
Contingency tables
 confidence
 origin
 standardization
Continuation-ratio logit model
 survival model
Continuity correction
Continuous proportions
Correlation
 exchangeable
 models
 predictive power
 test
 working, clustered data
Correspondence analysis
Cramér's V
Credible interval
Credit scoring
Cross-classification table, *see* Contingency tables
Cross-product ratio, *see* Odds ratio
Cross-validation
 k -fold
Cumulant function
Cumulative link model
 Bayesian fitting
 dispersion effects
Cumulative logit model
 Bayesian fitting
 conditional independence
 matched pairs
 matched sets
 random effects
 references
 sample size and power
Cumulative odds ratio
 uniform association
Cumulative probit model
 Bayesian fitting
 references
Data mining
Decomposable model
 existence of ML estimates
 graphical models contain
Degrees of freedom
 effect of estimating parameters
 moments of chi-squared
Delta method

Delta, two ordinal distributions

Dendrogram

Dependent proportions

clustered data

increased precision

matched pairs

Deviance

binomial GLMs

comparing models

goodness of fit

grouped vs. ungrouped binary data

information criterion (DIC)

Poisson GLM

residual

Dfbeta

Difference of proportions

Bayesian inference

chi-squared test

confidence interval

matched pairs

score confidence interval

score test

small-sample confidence interval

small-sample test

standard error

Differential item functioning

Dirichlet distribution

Dirichlet-multinomial distribution

Discordant pair

Discrete-choice model

ordered categories

Discreteness

complications

Discriminant analysis

diagonal

quadratic

vs. logistic regression

Dispersion effects

Dispersion parameter

Dissimilarity index

Dissimilarity, clustering

Diversity

Ecological diversity

Ecological inference

Effect modifier

EM algorithm

Empirical Bayes

Empirical logit

- Entropy
- Equiprobability
- Exact inference
 - conditional
 - conditional vs. unconditional
 - logistic regression
 - testing conditional independence
 - testing independence
 - unconditional
- Exponential dispersion family
 - multivariate
- Exponential family
- Extreme value distribution cdf for log-log link
 - latent for complementary log-log
 - utility and discrete choice
 - utility and multinomial logit
 - utility for logistic model
- False discovery rate
- Firth penalized likelihood
- Fisher scoring
- Fisher's exact test
 - Bayesian comparison
- Fisher, R. A., contributions to CDA
- Freeman–Tukey statistic
- Fuzzy inference
- Gambler's ruin problem
- Gamma (ordinal measure)
 - inference
 - Yule's Q
- Gauss–Hermite quadrature
- GEE, *see* Generalized estimating equations
- GEE2
- Generalized additive model
 - multinomial response
- Generalized CMH statistic
- Generalized estimating equations
 - GEE2
 - working correlations
- Generalized linear mixed models
 - Bayesian
- Generalized linear models
 - binary data
 - count data
 - covariance matrix
 - likelihood equations
 - link function
 - multivariate

- quasi-likelihood
- random component
- sufficient statistics
- systematic component

Generalized loglinear model

- marginal models

Genomic applications

Geometric distribution

Gibbs sampling

- cumulative probit model
- multinomial probit model

GLMM, *see* Generalized linear mixed models

Global odds ratio

Goodman and Kruskal lambda

Goodman and Kruskal tau

Goodman, Leo, contributions to CDA

Goodness of fit

- deviance

- Hosmer–Lemeshow test
- likelihood-ratio statistic
- likelihood-ratio test
- logistic model
- loglinear model
- Pearson statistic

Graphical models

Grouped data

- deviance different from ungrouped
- grouped vs. ungrouped binary data

Hat matrix

Hazard function

Hessian matrix

Heterogeneity

- multicenter clinical trials

Hierarchical Bayes

Hierarchical models

High-dimensional data

Highest posterior density intervals

Homogeneity

- odds ratios

Homogeneous association

- linear-by-linear
- loglinear model
- small-sample test
- symmetric property

Hosmer–Lemeshow test

Hypergeometric distribution

- mean and variance
- multivariate

noncentral

Incomplete table

Independence

2×2 tables

Bayesian testing

chi-squared tests

conditional

joint

loglinear model

marginal homogeneity

marginal vs. conditional

mutual

plus marginal homogeneity

score statistic

Independence from irrelevant alternatives

Indicator variables

Infinite estimates

finite with penalized likelihood

logistic regression

loglinear models

Influence diagnostics

Information matrix

observed vs. expected

Interaction

loglinear three-factor

loglinear two-factor

no three-factor

odds ratio definition

Interval variable

Intraclass correlation

Isotropic table

Item response models

Iterative proportional fitting

Iterative reweighted least squares

Jaccard index

Jeffreys prior distribution

binary regression

binomial parameter

comparing proportions

Firth penalized likelihood

multinomial parameters

Joint independence

Joint response models

Kappa agreement measure

Kendall's tau

Kernel smoothing

multinomial data

Kullback–Leibler distance

Lagrange multiplier test, *see* Score test

Laplace approximation

Large-sample distribution theory

likelihood-ratio statistic

logit/loglinear models

model parameter estimates

non-normal

Pearson statistic

probability estimates

residuals

Lasso

Latent class models

Latent variable

Bayesian modeling

hierarchical model

latent class models

multinomial models

probit model

proportional odds structure

LD₅₀

Leverage

Likelihood-ratio confidence interval

Likelihood-ratio test

comparing deviances

comparing models

independence in two-way table

theoretical justification

Linear Logit model

efficiency and scoring

log likelihood

score test

small-sample conditional inference

Linear probability model

likelihood equations

Linear trend

two-way table

Linear-by-linear association model

heterogeneous

homogeneous

isotropy

Link function

t inverse cdf

canonical

complementary log-log

generalized

identity

inverse cdf

log
log-log
logit
probit

Local odds ratio

large-sample distribution
uniform association

Logistic distribution

t distribution approximation

Logistic regression

2×2 tables
adjacent-categories logits
autoregressive structure
baseline-category logit
Bayesian fitting
case-control studies
categorical predictors
collapsibility
conditional
covariance matrix
design
diagnostics
extreme-value utility
goodness of fit
history
imbalance of outcomes

implied by normal explanatory variables

infinite estimates

likelihood equations

linear logit model

loglinear model connection

marginal model

model fitting

model selection

nonparametric random effects

parameter interpretation

random effects and marginal model both logistic

random effects models

retrospective studies

small-sample inference

subject-specific

Logistic-normal model

Logit

bias

confidence interval

history

standard error

Logit models, *see* Logistic regression

Logit-normal distribution

Loglinear model

Bayesian fitting

collapsibility

complex sampling designs

conditional independence

count response data

covariance matrix

fitting

four-way tables

generalized

goodness-of-fit test

hierarchical

history

homogeneous association

independence in two-way table

inference

infinite estimates

joint independence

large-sample theory

likelihood equations

logistic model connection

model selection

multinomial

mutual independence

no three-factor interaction

parameter constraints

probability estimates

random effects

saturated

three-way tables

Loglinear models

large-sample theory

Lowess

Machine learning

Mantel–Haenszel effect estimates

Mantel, Nathan, contributions to CDA

Marginal homogeneity

T-way tables

binary data

implied by symmetry

independence

matched sets

quasi-symmetry connection

Marginal likelihood

Marginal models

approximate relation with random effects models

binary matched pairs

- GEE fitting
- ML fitting
- multiway tables
- nominal matched pairs
- ordinal matched pairs
- random effects models comparison
- square tables
- vs. transitional models, for matched pairs
- Marginal symmetry
- Marginal vs. conditional associations
- Marginally-specified model
- Market basket data
- Markov chain Monte Carlo
- Markov chains
- Matched pairs
 - subject-specific model
 - Bayesian inference
 - bivariate binary response
 - marginal model
 - McNemar test
 - random effects model
- McNemar's test
 - Cochran–Mantel–Haenszel test connection
 - crossover study
 - generalized
 - paired t test
- Measurement error
- Median effective level
- Meta-analysis
 - Bayesian
- Mid P -value
 - confidence intervals
 - Fisher's exact test
- Mid distribution function
- Midranks
- Minimax estimate
- Minimum chi-squared estimation
- Minimum discrimination information
- Misclassification error
- Missing at random
- Missing completely at random
- Missing data
 - clustered data
 - two-way tables
- Mixed logit model
- Mixed-membership model
- Mixture models
 - beta-binomial

- latent class
- logistic-normal
- negative binomial
- nonparametric random effects
- Rasch mixture

- Model misspecification

- GEE methods

- Model selection

- logistic models

- loglinear models

- Model smoothing

- Monte Carlo methods

- Mosaic plot

- Multicollinearity

- Multilevel models

- Multinomial distribution

- Bayesian inference

- correlation structure

- inference

- likelihood function

- likelihood-ratio statistic

- multiple comparisons

- Pearson goodness-of-fit statistic

- Poisson connection

- properties

- Multinomial logit models

- Multinomial Poisson homogenous model

- Multinomial probit model

- Bayesian fitting

- discrete choice

- Multinomial sampling

- independent

- product

- Multiple comparisons

- Bonferroni method

- false discovery rate

- loglinear models

- multinomial parameters

- odds ratios

- proportions

- Multiple correspondence analysis

- Multiple imputation

- Multivariate hypergeometric distribution

- Mutual independence

- Natural exponential family

- Nearest neighbors

- classification

- smoothing

Negative binomial distribution
exponential family
mode
no. successes before k failures
Poisson connection
variance proportional to mean

Negative binomial GLMs

Nested logit model

Newton–Raphson method
logistic regression
loglinear model

Neyman modified chi-squared

No three-factor interaction

Nominal variable
conditional independence tests
modeling

Noncentral chi-squared distribution
noncentrality

O, o rates of convergence

Observational study

Occupational mobility

Odds ratio
 $I \times J$ tables
asymptotic distribution
Bayesian inference
confidence interval
global
history
homogeneity
local
logistic regression
Mantel–Haenszel estimate
properties
relative risk approximation
small-sample confidence interval
standard error of log
working association, GEE
Yule's Q connection

Offset

Ordered logit model

Ordered probit model

Ordinal data
adjacent-categories logit model
comparing two distributions
concordant and discordant pairs
conditional independence tests
continuation-ratio logit model
cumulative link models

- cumulative logit models
- cumulative probit model
- discrete-choice models
- independence test
- loglinear models
- power advantage
- trend test
- Ordinal quasi-symmetry
- Ordinal response
- Ordinal variable
- Ordinary least squares
 - linear discriminant analysis
 - linear probability model
 - ordinal response
- Outlier
- Overdispersion
 - beta-binomial models
 - binomial GLMs
 - impossible with Bernoulli
 - multinomial GLM
 - Poisson GLM
- P*-value
 - mid, *see* Mid *P*-value
 - randomized
 - two-sided
- Parameter constraints
- Parsimony
 - estimating proportions
 - model selection
 - model smoothing
- Partial tables
- Partitioning chi-squared
 - combining categories
 - comparing measures
 - loglinear models
 - trend test in $I \times 2$ table
- Pattern mixture models
- Pearson chi-squared statistic
 - $I \times 2$ table
 - 2×2 tables
 - comparing models
 - goodness of fit
 - grouped vs ungrouped binary data
 - independence test
 - moments
 - theoretical justification
- Pearson residual
 - binary regression

- binomial
- GLM
- Poisson GLM
- Pearson, Karl, contributions to CDA
- Pearson–Yule association controversy
- Penalized likelihood
- Penalized quasi-likelihood
- Perfect discrimination
- Perfect table
- Phi-squared
- Plus four confidence interval
- Poisson distribution
 - exponential family form
 - moment generating function
 - multinomial connection
 - negative binomial connection
 - overdispersion
 - properties
 - variance test
- Poisson GLM
 - common mean model
 - deviance
 - overdispersion
 - Pearson residual
 - random effects models
 - standardized residuals
- Poisson loglinear model
 - contingency tables
 - covariance matrix
 - likelihood equations
 - multinomial connection
- Polychoric correlation
- Population-averaged effect
- Positive likelihood-ratio dependence
- Positive predictive value
- Posterior interval
- Power
 - chi-squared tests
 - comparing proportions
- Power divergence statistic
- Predictive power
 - binary regression
 - linear discriminant analysis
 - ordinal models
- Prior distribution
 - beta
 - binary response probabilities
 - comparing proportions

- conjugate
- data augmentation
- Dirichlet
- improper
- multivariate normal

Probit link function

Probit model

- Bayesian fitting
- history
- interpreting effects
- likelihood equations
- multinomial
- ordered
- threshold model
- utility functions

Profile likelihood confidence interval

- capture–recapture
- difference of proportions
- odds ratio
- software

Propensity score

Proportional hazards model

Proportional odds model

- adjacent-categories logit
- cumulative logit
- testing fit

Proportional reduction in variation

- deviance

Proportions

- confidence intervals
- continuous
- difference
- ratio, *see* Relative risk

Quasi variances

Quasi-complete separation

Quasi-independence

- agreement modeling

Quasi-likelihood methods

- binomial overdispersion

- clustered data

- Poisson overdispersion

Quasi-symmetry model

- agreement modeling

- Bradley–Terry model

- collapsing

- marginal homogeneity test

- matched sets

- nonparametric logistic connection

references

square tables

R (software), text website

R-squared measures

Raking contingency table

Random effects models

autocorrelated random effects

binary data

binary matched pairs

count data

interpretations

marginal model comparison

misspecification

multilevel

multinomial

non-normal random effect

nonnegative marginal correlations

nonparametric

parameterizations

predicted random effects

probit link

Random forest

Random intercept model

Randomized test

Ranking outcome categories

Rasch mixture model

Rasch model

Rate data

Rater agreement

RC model

Bayesian

isotropy

Regressive logistic model

Regularization methods

Relative risk

attributable risk connection

Bayesian inference

confidence interval

odds ratio approximation

standard error

Residuals

contingency table

deviance

GLMs

loglinear models

Pearson

references

standardized

Retrospective studies
logistic regression
Ridge regression
Ridits
ROC curve
ordinal response
Row effects loglinear model
isotropy

S-PLUS
Sample proportion
admissible estimator
binomial parameter inference
minimax estimate
ML estimate

Sample size
comparing proportions
logistic regression
power

Sampling zero
Sandwich covariance matrix
SAS, text website
Saturated model
loglinear

Score confidence interval
difference of proportions
odds ratio
proportion
references
relative risk

Score test
binary regression
CMH test
comparing GLMs
difference of proportions
generalized CMH test
goodness of fit of GLM
linear logit model
multinomial models
Pearson chi-squared statistic
proportion
proportional odds and Wilcoxon

Scores
choice of

Selection models

Sensitivity
Sequential logit model, *see* Continuation-ratio logit model
Simpson's paradox
Simultaneous testing

Small-area estimation

Small-sample distribution theory

Smoothing

- binary data

- generalized additive model

- kernel

- model

- penalized likelihood

Software

- examples, www.stat.ufl.edu/~aa/cda/cda.html

- website for book

Sparse contingency table

- asymptotics

- CMH test

- smoothing

Spatial data

Specificity

Spline function

SPSS, text website

Square tables

Standardized regression coefficients

Standardized residuals

- 2×2 tables

- binomial

- GLM

- independence

- loglinear models

- score statistic for outlier

Stata, text website

Statistical vs. practical significance

StatXact

Stepwise procedures

Stereotype model

Stochastic ordering

- Bayesian evaluation

- Bayesian probability estimate

- discrete cdf and uniform

- ordinal response

- two cdfs

Stochastic process

Structural zero

Subject-specific effect

- binary matched pairs

- binary matched set

- matched set

- random effects models

Subject-specific table

Summarizing measures

Supervised vs. unsupervised learning

Support vector machines

Suppressor variable

Survival model

Symmetry

binary data

logistic/loglinear models

matched sets

square tables

t distribution approximation of logistic

Tetrachoric correlation

Threshold model

nonconstant variability

ordinal response

Time series

Tolerance distribution

Transition probability matrix

Transitional models

Tree-structured classification

Trend tests

Uncertainty coefficient

and G^2

Uniform association model

cumulative odds ratio

global odds ratio

local odds ratio

Uniform interaction model

Unsupervised vs. supervised learning

Upper-triangular table

Utility

logistic model

multinomial probit model

probit model

Variance stabilizing transformations

Wald statistic

aberrant behavior for logistic regression

confidence interval

dependence on parameterization

infinite estimate

Weight matrix

Weighted kappa

Weighted least squares

Wilcoxon test

cumulative logit model

Within-subject effects

Yule's Q

asymptotic variance

Yule, G. U., contributions to CDA

Zero count

effect

infinite estimates

odds ratio

sampling zero

structural

Zero-inflated models