

Log-Linear Models and Logistic Regression

Ronald Christensen
Department of Mathematics and Statistics
University of New Mexico

To Sharon and Fletch

Preface to the Second Edition

As the new title indicates, this second edition of *Log-Linear Models* has been modified to place greater emphasis on logistic regression. In addition to new material, the book has been radically rearranged. The fundamental material is contained in Chapters 1-4. Intermediate topics are presented in Chapters 5 through 8. Generalized linear models are presented in Chapter 9. The matrix approach to log-linear models and logistic regression is presented in Chapters 10-12, with Chapters 10 and 11 at the applied Ph.D. level and Chapter 12 doing theory at the Ph.D. level.

The largest single addition to the book is Chapter 13 on Bayesian binomial regression. This chapter includes not only logistic regression but also probit and complementary log-log regression. With the simplicity of the Bayesian approach and the ability to do (almost) exact small sample statistical inference, I personally find it hard to justify doing traditional large sample inferences. (Another possibility is to do exact conditional inference, but that is another story.)

Naturally, I have cleaned up the minor flaws in the text that I have found. All examples, theorems, proofs, lemmas, etc. are numbered consecutively within each section with no distinctions between them, thus Example 2.3.1 will come before Proposition 2.3.2. Exercises that do not appear in a section at the end have a separate numbering scheme. Within the section in which it appears, an equation is numbered with a single value, e.g., equation (1). When reference is made to an equation that appears in a different section, the reference includes the appropriate chapter and section, e.g., equation (2.1.1).

The primary prerequisite for using this book is knowledge of analysis

of variance and regression at the masters degree level. It would also be advantageous to have some prior familiarity with the analysis of two-way tables of count data. Christensen (1996a) was written with the idea of preparing people for this book and for Christensen (1996b). In addition, familiarity with masters level probability and mathematical statistics would be helpful, especially for the later chapters. Sections 9.3, 10.2, 11.6, and 12.3 use ideas of the convergence of random variables. Chapter 12 was originally the last chapter in my linear models book, so I would recommend a good course in linear models before attempting that. A good course in linear models would also help for Chapters 10 and 11.

The analysis of logistic regression and log-linear models is not possible without modern computing. While it certainly is not the goal of this book to provide training in the use of various software packages, some examples of software commands have been included. These focus primarily on SAS and BMDP, but include some GLIM (of which I am still very fond).

I would particularly like to thank Ed Bedrick for his help in preparing this edition and Ed and Wes Johnson for our collaboration in developing the material in Chapter 13. I would also like to thank Turner Ostler for providing the trauma data and his prior opinions about it.

Most of the data, and all of the larger data sets, are available from STATLIB as well as by anonymous ftp. The web address for the datasets option in STATLIB is <http://www.stat.cmu.edu/datasets/>. The data are identified as "christensen-llm". To use ftp, type **ftp stat.unm.edu** and login as "anonymous", enter **cd /pub/fletcher** and either get **llm.tar.Z** for Unix machines or **llm.zip** for a DOS version. More information is available from the file "readme.llm" or at <http://stat.unm.edu/~fletcher>, my web homepage.

Ronald Christensen
Albuquerque, New Mexico
February, 1997

BMDP Statistical Software is distributed by SPSS Inc., 444 N. Michigan Avenue, Chicago, IL, 60611, telephone: (800) 543-2185.

MINITAB is a registered trademark of Minitab, Inc., 3081 Enterprise Drive, State College, PA 16801, telephone: (814) 238-3280, telex: 881612.

MSUSTAT is marketed by the Research and Development Institute Inc., Montana State University, Bozeman, MT 59717-0002, Attn: R.E. Lund.

Preface to the First Edition

This book examines log-linear models for contingency tables. Logistic regression and logistic discrimination are treated as special cases and generalized linear models (in the GLIM sense) are also discussed. The book is designed to fill a niche between basic introductory books such as Fienberg (1980) and Everitt (1977) and advanced books such as Bishop, Fienberg, and Holland (1975), Haberman (1974a), and Santner and Duffy (1989). It is primarily directed at advanced Masters degree students in Statistics but it can be used at both higher and lower levels. The primary theme of the book is using previous knowledge of analysis of variance and regression to motivate and explicate the use of log-linear models. Of course, both the analogies and the distinctions between the different methods must be kept in mind.

[From the first edition, Chapters I, II, and III are about the same as the new 1, 2, and 3. Chapter IV is now Chapters 5 and 6. Chapter V is now 7, VI is 10, VII is 4 (and the sections are rearranged), VIII is 11, IX is 8, X is 9, and XV is 12.]

The book is written at several levels. A basic introductory course would take material from Chapters I, II (deemphasizing Section II.4), III, Sections IV.1 through IV.5 (eliminating the material on graphical models), Section IV.10, Chapter VII, and Chapter IX. The advanced modeling material at the end of Sections VII.1, VII.2, and possibly the material in Section IX.2 should be deleted in a basic introductory course. For Masters degree students in Statistics, all the material in Chapters I through V, VII, IX, and X should be accessible. For an applied Ph.D. course or for advanced Masters students, the material in Chapters VI and VIII can be incorporated. Chapter VI recapitulates material from the first five chapters using matrix notation. Chapter VIII recapitulates Chapter VII. This material is necessary (a) to get standard errors of estimates in anything other than the saturated model, (b) to explain the Newton-Raphson (iteratively reweighted least squares) algorithm, and (c) to discuss the weighted least

squares approach of Grizzle, Starmer, and Koch (1969). I also think that the more general approach used in these chapters provides a deeper understanding of the subject. Most of the material in Chapters VI and VIII requires no more sophistication than matrix arithmetic and being able to understand the definition of a column space. All of the material should be accessible to people who have had a course in linear models. Throughout the book, Chapter XV of Christensen (1987) is referenced for technical details. For completeness, and to allow the book to be used in nonapplied Ph.D. courses, Chapter XV has been reprinted in this volume under the same title, Chapter XV.

The prerequisites differ for the various courses described above. At a minimum, readers should have had a traditional course in statistical methods. To understand the vast majority of the book, courses in regression, analysis of variance, and basic statistical theory are recommended. To fully appreciate the book, it would help to already know linear model theory.

It is difficult for me to understand but many of my acquaintance view me as quite opinionated. While I admit that I have not tried to keep my opinions to myself, I have tried to clearly acknowledge them as my opinions.

There are many people I would like to thank in connection with this work. My family, Sharon and Fletch, were supportive throughout. Jackie Damrau did an exceptional job of typing the first draft. The folks at BMDP provided me with copies of 4F, LR, and 9R. MINITAB provided me with Versions 6.1 and 6.2. Dick Lund gave me a copy of MSUSTAT. All of the computations were performed with this software or GLIM. Several people made valuable comments on the manuscript; these include Rahman Azari, Larry Blackwood, Ron Schrader, and Elizabeth Slate. Joe Hill introduced me to statistical applications of graph theory and convinced me of their importance and elegance. He also commented on part of the book. My editors, Steve Fienberg and Ingram Olkin, were, as always, very helpful. Like many people, I originally learned about log-linear models from Steve's book. Two people deserve special mention for how much they contributed to this effort. I would not be the author of this book were it not for the amount of support provided in its development by Ed Bedrick and Wes Johnson. Wes provided much of the data used in the examples. I suppose that I should also thank the legislature of the state of Montana. It was their penury, while I worked at Montana State University, that motivated me to begin the project in the spring of 1987. If you don't like the book, blame them!

Ronald Christensen
Albuquerque, New Mexico
April 5, 1990
(Happy Birthday Dad)

Contents

Preface to the Second Edition	vii
Preface to the First Edition	ix
1 Introduction	1
1.1 Conditional Probability and Independence	2
1.2 Random Variables and Expectations	11
1.3 The Binomial Distribution	13
1.4 The Multinomial Distribution	14
1.5 The Poisson Distribution	18
1.6 Exercises	21
2 Two-Dimensional Tables and Simple Logistic Regression	24
2.1 Two Independent Binomials	24
2.1.1 The Odds Ratio	30
2.2 Testing Independence in a 2×2 Table	31
2.2.1 The Odds Ratio	33
2.3 $I \times J$ Tables	34
2.3.1 Response Factors	38
2.3.2 Odds Ratios	39
2.4 Maximum Likelihood Theory for Two-Dimensional Tables .	43
2.5 Log-Linear Models for Two-Dimensional Tables	48
2.5.1 Odds Ratios	52
2.6 Simple Logistic Regression	55

2.6.1	Computer Commands	62
2.7	Exercises	62
3	Three-Dimensional Tables	70
3.1	Simpson's Paradox and the Need for Higher-Dimensional Tables	71
3.2	Independence and Odds Ratio Models	73
3.2.1	The Model of Complete Independence	73
3.2.2	Models with One Factor Independent of the Other Two	76
3.2.3	Models of Conditional Independence	80
3.2.4	A Final Model for Three-Way Tables	84
3.2.5	Odds Ratios and Independence Models	86
3.3	Iterative Computation of Estimates	88
3.4	Log-Linear Models for Three-Dimensional Tables	90
3.4.1	Estimation	93
3.4.2	Testing Models	95
3.5	Product-Multinomial and Other Sampling Plans	100
3.5.1	Other Sampling Models	103
3.6	Model Selection Criteria	105
3.6.1	R^2	105
3.6.2	Adjusted R^2	106
3.6.3	Akaike's Information Criterion	107
3.7	Higher-Dimensional Tables	109
3.7.1	Computer Commands	111
3.8	Exercises	114
4	Logistic Regression, Logit Models, and Logistic Discrimination	118
4.1	Multiple Logistic Regression	122
4.1.1	Informal Model Selection	124
4.2	Measuring Model Fit	129
4.2.1	Checking Lack of Fit	131
4.3	Logistic Regression Diagnostics	132
4.4	Model Selection Methods	138
4.4.1	Computations for Nonbinary Data	140
4.4.2	Computer Commands	141
4.5	ANOVA Type Logit Models	143
4.5.1	Computer Commands	151
4.6	Logit Models for a Multinomial Response	152
4.7	Logistic Discrimination and Allocation	161
4.8	Exercises	171
5	Independence Relationships and Graphical Models	179
5.1	Model Interpretations	179
5.2	Graphical and Decomposable Models	182

5.3	Collapsing Tables	193
5.4	Recursive Causal Models	195
5.5	Exercises	210
6	Model Selection Methods and Model Evaluation	212
6.1	Stepwise Procedures for Model Selection	213
6.2	Initial Models for Selection Methods	216
6.2.1	All s -Factor Effects	216
6.2.2	Examining Each Term Individually	218
6.2.3	Tests of Marginal and Partial Association	218
6.2.4	Testing Each Term Last	219
6.3	Example of Stepwise Methods	225
6.3.1	Forward Selection	227
6.3.2	Backward Elimination	231
6.3.3	Comparison of Stepwise Methods	233
6.3.4	Computer Commands	234
6.4	Aitkin's Method of Backward Selection	235
6.5	Model Selection Among Decomposable and Graphical Models	241
6.6	Use of Model Selection Criteria	246
6.7	Residuals and Influential Observations	248
6.7.1	Computations	250
6.7.2	Computing Commands	254
6.8	Drawing Conclusions	255
6.9	Exercises	257
7	Models for Factors with Quantitative Levels	259
7.1	Models for Two-Factor Tables	260
7.1.1	Log-Linear Models with Two Quantitative Factors	261
7.1.2	Models with One Quantitative Factor	263
7.2	Higher-Dimensional Tables	267
7.2.1	Computing Commands	269
7.3	Unknown Factor Scores	270
7.4	Logit Models	275
7.5	Exercises	278
8	Fixed and Random Zeros	280
8.1	Fixed Zeros	280
8.2	Partitioning Polytomous Variables	283
8.3	Random Zeros	287
8.4	Exercises	294
9	Generalized Linear Models	298
9.1	Distributions for Generalized Linear Models	300
9.2	Estimation of Linear Parameters	305
9.3	Estimation of Dispersion and Model Fitting	307

9.4	Summary and Discussion	312
9.5	Exercises	314
10	The Matrix Approach to Log-Linear Models	315
10.1	Maximum Likelihood Theory for Multinomial Sampling . .	319
10.2	Asymptotic Results	323
10.3	Product-Multinomial Sampling	340
10.4	Inference for Model Parameters	343
10.5	Methods for Finding Maximum Likelihood Estimates . . .	346
10.6	Regression Analysis of Categorical Data	348
10.7	Residual Analysis and Outliers	355
10.8	Exercises	361
11	The Matrix Approach to Logit Models	364
11.1	Estimation and Testing for Logistic Models	364
11.2	Model Selection Criteria for Logistic Regression	372
11.3	Likelihood Equations and Newton-Raphson	373
11.4	Weighted Least Squares for Logit Models	376
11.5	Multinomial Response Models	377
11.6	Asymptotic Results	379
11.7	Discrimination, Allocation, and Retrospective Data	387
11.8	Exercises	394
12	Maximum Likelihood Theory for Log-Linear Models	396
12.1	Notation	396
12.2	Fixed Sample Size Properties	397
12.3	Asymptotic Properties	402
12.4	Applications	412
12.5	Proofs of Lemma 12.3.2 and Theorem 12.3.8	418
13	Bayesian Binomial Regression	422
13.1	Introduction	422
13.2	Bayesian Inference	424
13.2.1	Specifying the Prior and Approximating the Posterior	424
13.2.2	Predictive Probabilities	434
13.2.3	Inference for Regression Coefficients	436
13.2.4	Inference for LD_α	438
13.3	Diagnostics	441
13.3.1	Case Deletion Influence Measures	441
13.3.2	Model Checking	446
13.3.3	Link Selection	447
13.3.4	Sensitivity Analysis	448
13.4	Posterior Computations and Sample Size Calculation	449
	Appendix: Tables	455

A.1 The Greek Alphabet	455
A.2 Tables of the χ^2 Distribution	456
References	458
Author Index	472
Subject Index	476