

ANALYZING QUALITATIVE DATA

Log-Linear Analysis for
Behavioral Research
Second Edition

JOHN J. KENNEDY

THE ORIGIN OF VALUES

Michael Hechter, Lynn Nadel, and
Richard E. Michod

EDITORS



ALDINE DE GRUYTER
New York

Copyright Acknowledgments

The author and publisher gratefully acknowledge permission to use the following copyrighted materials.

Goldsmid, C. A., Gruber, J. E., and Wilson, E. K. "Perceived Attributes of Superior Teachers (PAST): An Inquiry into the Giving of Teacher Awards." *American Educational Research Journal*, 14 (1977), 423-440. Copyright © 1977 by the American Educational Research Association, Washington, D.C. Reprinted by permission.

O'Connor, G., and Sitkei, E. G. "Study of a New Frontier in Community Services: Residential Facilities for the Developmentally Disabled." *Mental Retardation*, 13, 4 (1975), 35-39. Copyright © 1975 by the American Association on Mental Deficiency. Reprinted by permission.

Library of Congress Cataloging-in-Publication Data

Kennedy, John J.

Analyzing qualitative data : log-linear analysis for behavioral
research / John J. Kennedy. — 2nd ed.

p. cm.

Includes bibliographical references and indexes.

ISBN 0-275-93446-2 (alk. paper)

1. Log-linear models. 2. Psychometrics. 3. Psychology—Research.
4. Education—Research. I. Title.
BF39.K457 1992
519.5'02415—dc20 91-4936

British Library Cataloguing in Publication Data is available.

Copyright © 1992 by John J. Kennedy

All rights reserved. No portion of this book may be
reproduced, by any process or technique, without the
express written consent of the publisher.

Library of Congress Catalog Card Number: 91-4936
ISBN: 0-275-93446-2

First published in 1992

Praeger Publishers, One Madison Avenue, New York, NY 10010
An imprint of Greenwood Publishing Group, Inc.

Printed in the United States of America



The paper used in this book complies with the
Permanent Paper Standard issued by the National
Information Standards Organization (Z39.48-1984).

10 9 8 7 6 5 4 3 2 1

Preface

The first edition of this book was published in 1983 under the title *Analyzing Qualitative Data: Introductory Log-Linear Analysis for Behavioral Research*. Its purpose was to illuminate the basic features of log-linear analysis. At that time, the majority of researchers in behavioral fields knew very little about this relatively new generalized approach to the analysis of contingency table data. The use and acceptance of the 1983 edition demonstrated that a book-length introductory treatment of log-linear modeling was both needed and appreciated.

During the 1980s, the use of log-linear statistical models in behavioral and life-science inquiry has increased markedly. Concurrently, log-linear theory, which was developed largely during the previous decade, has been streamlined and refined. An aim of this second edition is to acquaint old and new readers with these refinements.

The biggest change that occurred during the 1980s, however, has been the increased availability of user-oriented computer programs for the performance of log-linear analyses. During this period, all major statistical packages (i.e., BMDP, SAS, and SPSS) introduced either new or improved computer programs designed specifically for the specification and fitting of log-linear models. As a consequent, the enhanced ability of practicing researchers to perform log-linear analyses has been accompanied by an enhanced need for didactic explanations of this system of analysis—for explanations of log-linear theory and method that can be readily understood by practitioners and graduate students who do not possess recondite backgrounds in mathematical statistics yet desire to obtain a level of understanding beyond that which is typically offered by cookbook approaches to statistical topics. Another aim of this second edition is to fulfill this need.

As before, this book has been prepared for readers who have had at least one

intermediate-level course in applied statistics in which the basic principles of factorial analysis of variance and multiple regression were discussed. Also as before, to assist readers with modest preparation in the analysis of quantitative/categorical data, this edition will review topics in such relevant areas as basic probability theory, traditional chi-square goodness-of-fit procedures, and the method of maximum-likelihood estimation. Readers with strong backgrounds in statistics can skim over these preparatory discussions, which are contained largely in Chapters 2 and 3, without prejudice.

Because the origins of log-linear methodology are diverse, several approaches can be taken to its formulation and application. The general approach of the 1983 edition, maintained in this second edition, is associated with Professor Leo A. Goodman of the University of Chicago. Readers who have been exposed to the Goodman approach, however, will occasionally find treatments of topics in this book that are somewhat at variance with Goodman's writings and the writings of his associates. Aside from the modest mathematical requisites, experimental and ex post facto research rather than the survey are emphasized. Expect, therefore, to see relatively greater emphasis given here to research situations in which the intent is to determine whether groups of subjects differ with respect to a categorical response (or dependent) variable. In fact, beginning with Chapter 6, attention will be directed almost exclusively to the study of group differences with respect to one or more response variables. As will be shown, special log-linear models, known as asymmetrical or logit models, will be used to estimate group and interaction effects in a manner that resembles the estimation and testing of effects in the analysis of variance. Moreover, interpretation of group effects is more apt to be couched in terms of differences between geometric means (on a log scale) as opposed to odds ratios. In sum, the approach taken to log-linear modeling in this book is essentially that advanced by Goodman, but I have modified it to exploit, whenever possible, the synonymity between log-linear analysis and the analysis of variance. As suggested by the success of the first edition, this modification has been appreciated especially among researchers in fields with strong experimental traditions, such as psychology and medicine.

Finally, readers familiar with the 1983 edition will find that in addition to expanded treatments of numerous topics, two entirely new topics are introduced in the present edition. There is a new chapter (Chapter 6) devoted to a discussion of follow-up procedures, termed *focused comparisons*, that resemble in a number of respects *t* test contrasts that are often computed to clarify the results following an analysis of variance. Also new is a chapter (Chapter 9) devoted to an introductory discussion of *configural frequency analysis*, a taxonomical technique rooted in log-linear methodology that may be used to identify discernable groups of subjects who share a common profile of response to multiple categorical variables. I trust that both old and new readers of this book will find these new topics to be of immense value.

Finally, I am deeply grateful to a number of people who, directly or indirectly,

helped me with this edition. To Dr. Andrew J. Bush of Baptist Memorial Hospital in Memphis and Dr. Janet C. Rice in the Department of Biostatistics and Epidemiology at Tulane University, two valued colleagues with whom I have had the opportunity to offer a number of short professional courses on log-linear methods, go my special thanks for their friendship, encouragement, and intellectual guidance. I am similarly indebted to one of my current and best students, Hak Ping Tam, and to one of my former students, Debora Ann Grale of North Coast Behavioral Research in Cleveland, for their long hours, insights, and special help with the preparation of this manuscript. In addition, for her help with the manuscript, my thanks go to my secretary, Barbara E. Heinlein.

Contents

	Figures and Tables	xi
	Preface	xvii
1	Qualitative Data and Statistical Techniques	1
	Qualitative Variables and Data	2
	Basic Considerations	6
	The Forest Before the Trees	13
2	Statistical Foundations	19
	Dichotomies and the Binomial Law	20
	Polytomies and the Multinomial Law	31
	The Product-Multinomial Rule	34
	Maximum-Likelihood Estimation	36
3	Traditional Applications of Chi-Square	43
	The χ^2 and Its Sampling Distribution	43
	Applications and Requisite Conditions	45
	Single-Variable Goodness-of-Fit	48
	The Hypothesis of Variable Independence	50
	The Hypothesis of Homogeneity of Proportions	55
	The Likelihood-Ratio Chi-Square	59
	Residual Analysis and Other Follow-Up Procedures	62

4	Log-Linear Analysis of Two-Dimensional Tables	69
	Major Operations	69
	A First Look at Contingency Table Models	72
	Comparing Observed and Expected Frequencies	91
	An Initial Discussion of Model Selection	96
	Model Interpretation and Follow-Up Procedures	98
5	The Analysis of Three-Dimensional Tables	107
	Working Examples	107
	Hierarchical Models for Symmetrical Inquiry	111
	Interpreting General Models in Symmetrical Inquiry	113
	Model Selection in the Symmetrical Case	124
	Interpreting Effects in Asymmetrical Inquiry	131
	Assessing Effects in the Asymmetrical Case	137
	The Effects of Reflective Teaching: A Didactic Example	143
6	Following Up Asymmetrical Analyses	153
	A Working Example	154
	The Methodology of Focused Comparisons	157
	The Holton-Nott Study Revisited	179
	Concluding Comments	182
7	Tables of High Dimensionality	185
	The Symmetrical Analysis of Four-Dimensional Tables	185
	The Asymmetrical Analysis of Four-Dimensional Tables	196
	Explicating Interaction in a Logit-Model Analysis	203
	The Analysis of Five-Dimensional Tables	208
8	Additional Applications and Concerns	221
	Multidimensional Logits	222
	Repeated Observations and Response Independence	225
	Causal Model Analogues	237
	Additional Concerns	247
	The Need for Log-Linear Methods	251
9	An Introduction to Configural Frequency Analysis	257
	Basic Concepts and Nomenclature	258
	A Global Model Example	266
	An Asymmetrical Example	274
	Concluding Remarks	279

Appendix A: An SAS Program for Configural Frequency Analysis	283
Appendix B: Critical Values for Chi-Square Statistics	287
References	289
Name Index	295
Subject Index	297

Figures and Tables

FIGURES

2.1	Expected Relative Frequency of Observing a Six in Four Rolls of a Die	26
2.2	Lower-Tail Probabilities for a Binomial Sampling Distribution Where $n = 10$ and $P = .30$	28
3.1	Chi-Square Distributions Associated with 2, 10, and 30 Degrees of Freedom	45
6.1	Plot of Log Odds (from Table 6.8) Associated with FC_{22} , by Marital Status and Gender	175
7.1	Plot of Binomial Logits by Race of Student and Type of University	207
8.1	Path Diagram Showing Four Main Variables and an Interactive Variable to Be Used in Conjunction with the McLean Study	238
8.2	Revised Path Diagram for the McLean Study with Path Coefficients (Lambda Parameter Estimates) and z Tests (in Parentheses) Performed on the Coefficients	245

TABLES

1.1	Cross-Tabulations (Frequencies) by Gender and Attitude Toward a Constitutional Amendment: An Example of a 2×2 Contingency Table	10
-----	--	----

1.2	Taxonomy of Statistical Techniques for Qualitative Data by Mode of Inquiry and Nature of Defining Variables	14
2.1	Distinct Sequences and Their Probability of Occurrence for Four Bernoulli Trials	23
3.1	Observed and Expected Frequencies for the Hypothetical Study Introduced in Chapter 1	53
3.2	Cross-Classification of Facilities by Type and Number of Retarded Residents Served	54
3.3	Cross-Classifications by Type of Teacher and Age	57
3.4	Proportional Response to Levels of Age by Type of Teacher	58
3.5	Standardized Residuals (R_s) and Freeman-Tukey Deviates (D_s) as Applied to the O'Connor-Sitkei Data in Table 3.3	65
4.1	Hierarchical Multiplicative Models for Two-Dimensional Tables	90
4.2	Summary of Log-Linear Analysis of Data Presented in Table 1.1	95
4.3	Observed $[AB]$ Frequencies, Lambda Parameters, and Normal Deviate Tests for the Working Example	106
5.1	Observed Frequencies Associated with the Symmetric Substance Abuse Study and the Asymmetric Adverse Impact Study	108
5.2	Hierarchical Multiplicative Models for Three-Dimensional Tables	112
5.3	Hierarchical Log-Linear Models for Three-Dimensional Tables	114
5.4	Adequacy-of-Fit of General Log-Linear Models for the Substance Abuse Data	126
5.5	Summary Table of the Logit-Model Analysis of Adverse Impact Data	138
5.6	Cross-Classification of Subjects in the Reflective Teaching Experiment by Pretest Response, Treatment Groups, and Posttest Response	144
5.7	Residual Chi-Squares Computed on Logit-Models for the Reflective Teaching Experiment	145
5.8	Summary Table of the Logit-Model Analysis of the Reflective Teaching Experiment	146

5.9	A Comparison of Three Approaches to the Calculation, Interpretation, and Reporting of Effects in the Holton-Nott Study	148
5.10	Observed Frequencies, Lambdas, and Tests on Lambdas for a Follow-Up Model Fitted to the Holton-Nott Data	149
6.1	Observed Frequencies by Gender (Variable <i>A</i>), Marital Status (Variable <i>B</i>), and Response (Variable <i>C</i>)	155
6.2	Summary of Asymmetrical Analysis of Response (Variable <i>C</i>) by Gender (Variable <i>A</i>) and Marital Status (Variable <i>B</i>)	156
6.3	General Types of <i>FC</i> s Used to Explicate Main Effects for <i>B</i> and Interaction in the Working Example	159
6.4	Summary of Orthogonal Marginal Comparisons and the Partitioning of the Component L^2 for the <i>BC</i> Term	164
6.5	Expected Elementary Cell Frequencies Generated by the Full Main-Effects Model (Model 7)	167
6.6	Summary of Orthogonal Fully Partialled Comparisons Based on the Partitioning of the Partial Component for <i>BC</i>	169
6.7	Summary of Orthogonal Compound Comparisons and the Partitioning of the Component for BC and ABC	172
6.8	FC_{22} : Log Odds of Giving a Nonneutral Response as Opposed to a Neutral Response, by Marital Status and Gender	174
6.9	Summary of Orthogonal Interactive Comparisons and the Partitioning of the Component Chi-Square for <i>ABC</i>	178
6.10	Selected (Nonorthogonal) Partial <i>FC</i> s Applied to Results Reported by Holton and Nott	181
7.1	A Hierarchical Arrangement of Models for Tables of Four Dimensions	187
7.2	Screening Tests on Full Models Applied to the McLean Data	189
7.3	Residual and Component Chi-Squares for Selected Models Fitted to the McLean Data	190

7.4	Tests of Marginal and Full Partial Associations and Tests of Lambda Parameters for the McLean Data	193
7.5	Final Fit of General Models to the McLean Data	195
7.6	Hierarchical Logit Models for Four-Way Tables Where D Is the Logit Variable	197
7.7	The McLean Data: Observed Frequencies by Race, Type of University, Ability Level, and Completion Status	200
7.8	Summary of Two Approaches to the Logit-Model Analysis on the McLean Data	202
7.9	Estimates of Logit Interaction Effects Between Variables B and C in the McLean Data	204
7.10	Odds and Binomial Logits from Response Variable D as Observed in Configuration (BC) of the McLean Data	206
7.11	A Hierarchical Arrangement of Logit Models for Four Explanatory Variables	210
7.12	Summary of Logit-Model Analysis of MBTI Data	212
7.13	Estimates of Main Effects for the Dimensions of Thinking (T 's) Versus Feelings (F 's)	213
7.14	First-Order-Logit Interaction Effects for Variables B and C	216
8.1	Two-Dimensional Logit Models Where Variables A and B Are Explanatory	224
8.2	Hypothetical Data Sets for the Holton-Nott Study	227
8.3	Observed Frequencies and Those Expected Under the Hypothesis of Symmetrical Change	229
8.4	A Reorganization of Observed Frequencies and Those Expected Under the Hypothesis of Symmetrical Change	231
8.5	Summary of Analyses of Change of Treated Subjects in the Holton-Nott Study	236
8.6	Component L^2 's for Model Fitted to the McLean Data in the Performance of a Causal Analysis	242
9.1	Fit of Global k th-Order Models to the Lienert-Krauth Psychopathology Data	267
9.2	Global First-Order CFA of Lienert-Krauth Psychopathology Data	269

9.3	Global Second-Order CFA of Lienert-Krauth Psychopathology Data	271
9.4	Fit of Selected Log-Linear Models Applied to Assess the Underlying Structure of Second-Order Configurations	272
9.5	Global Second-Order CFA of Lienert-Krauth Psychopathology Data on Reduced $B \times T \times A$ Table	273
9.6	Prefatory Screening Analysis of Incipient Teacher Data	277
9.7	LMCFA of Incipient Teacher Data	278

Qualitative Data and Statistical Techniques

Qualitative data, also known as categorical data, are cross-tabulations within the context of contingency tables. Until recently, the analysis of qualitative data was limited, for the most part, to data within two-dimensional contingency tables. For such tables, the use of chi-square goodness-of-fit procedures, which were developed early in the history of statistics by Karl Pearson, have proven to be of immense value. Often, however, researchers possess information that is sufficient to structure contingency tables of more than two dimensions. Further, they often desire to subject tables of higher dimensionality to a comprehensive, simultaneous analysis. Unfortunately, prior to the 1970s, the methodology that would permit the desired analysis was not refined sufficiently to be of assistance to practicing researchers.

Then, due to the pioneering efforts of Grizzle, Starmer, and Koch (1969), Goodman (1970, 1971a, 1971b), Bishop, Fienberg, and Holland (1975), Bock (1975, ch. 8), and Krauth and Lienert (1973), among others, a revolution transpired in the analysis of multidimensional cross-tabular data. Major advances developed along at least three distinct lines.

The first line of development can be traced directly to Grizzle, Starmer, and Koch (1969). Their approach, sometimes called the GSK approach, employs the method of *weighted least-squares* (WLS) regression to estimate parameters in models that are formulated to explain categorical response variables. WLS regression, as used in the GSK approach, is particularly flexible with respect to analyzing response variables that consist of more than two categories.

A second line of development has been the application of log-linear models to qualitative data. Log-linear analysis has had numerous contributors. The most prominent contributions, however, emanate from a text authored by Bishop, Fienberg, and Holland (1975) and from the extensive writings of Goodman (e.g.,

1978). The Goodman approach in particular has been cited frequently in the behavioral research literature, especially in the sociological literature. Moreover, this approach has been explicated for practitioners in a variety of fields via a number of didactic papers (Shaffer, 1973; Marks, 1975; Knoke & Burke, 1980; Kennedy, 1982, 1988; Green, 1988).

In addition to more frequent usage, log-linear models differ from the WLS models of the GSK approach in two important respects. First, parameters that appear in log-linear models are estimated by the method of *maximum likelihood* (ML), not ordinary least-squares (OLS) or WLS. Second, log-linear models can be used either to investigate *relationships* between or among variables (i.e., a *symmetrical analysis*) or to determine the *effects* of selected categorical variables on a designated response variable (an *asymmetrical analysis*). WLS modeling, in contrast, is limited, for the most part, to the latter type of determination.

The third development occurred in West Germany where, during the 1970s, methodologists were exploring the use of regression and log-linear models (specifically the analysis of residuals from these models) as a means of identifying personality types and clinical syndromes (Krauth & Lienert, 1973). Their work has produced a methodology known as *configural frequency analysis* (CFA). Briefly, CFA is a classification technique that attempts to identify discernable clusters or groups of subjects that manifest similar profiles of response on two or more categorical variables.

As before, the principal intent of this second edition is to introduce behavioral and life-science researchers to the theory, operations, and typical applications of log-linear contingency table analysis, and to do so, whenever possible, within a conceptual framework that is compatible with the analysis of variance (AN-OVA). The contents of the initial edition have been updated and, in a number of areas, expanded, consonant with advancements that have appeared during the 1980s. Entirely new to this second edition, however, is a chapter on focused comparisons (Chapter 6), a methodology designed to assist analysts when they have a need to follow up the results of a log-linear analysis, and a chapter on configural frequency analysis (Chapter 9). It is hoped that the exposition of log-linear methodology, augmented with discussions of follow-up procedures and CFA, will increase still further the number of practitioners who will use these relatively new techniques in their research and appreciate their use in the work of others.

QUALITATIVE VARIABLES AND DATA

Ironically, terms most basic to an area of study often prove to be most difficult to define. Definitions are required, nevertheless, and because this volume deals with the study of categorical variables and resultant qualitative data, an attempt will be made to clarify these terms. We will attend first to a discussion of the general nature of data, returning subsequently to a brief discussion of categorical variables.

The Nature of Data

In most instances, data gathered during the course of a behavioral investigation consists of *measurements*, numbers assigned to objects or events according to a set of rules. Implied by this definition of measurements is that there are a number of rules used in the process of assigning numbers. Due largely to the scholarship of Stevens (1946), at present there is substantial agreement that the process of number assignment (i.e., measurement) is governed by four fundamental sets of rules. These four sets of rules give rise to four *scales of measurement*, or, more simply, four types of data: nominal, ordinal, interval, and ratio data. Since these four types of data are discussed extensively in most introductory textbooks in statistics and measurement, only the principal features of each are reviewed below.

Nominal Data. Nominal scaling connotes classification. Briefly, the objects or events of concern are examined for underlying similarities (or differences) and subsequently grouped on the basis of observed *qualitative* distinctions. Examples of nominal scaling abound. For instance, nominal scaling is implemented when research subjects are classified according to their gender, sexual orientation, ethnic background, type of learning disability, or similar characteristics. Now, if the numbers of subjects falling into respective categories of a variable such as gender or ethnic background are tallied, the resultant counts or frequencies constitute nominal data. Simply put, nominal data consists of frequencies observed within categories of a nominal or categorical variable.

Notice that nominal scaling barely conforms to the definition of measurement presented earlier since in most instances the act of assigning numbers is optional; moreover, when numbers are assigned to groups or categories, they serve merely as labels to facilitate coding or identification. Because qualitative categorization is so rudimentary and the act of assigning numbers is so arbitrary, many theorists do not seriously regard nominal scaling as a formal measurement process. Nominal scaling merits serious consideration, however, for not only is this manner of measurement frequently used in behavioral research, but the analysis of nominal data also constitutes the subject matter of this text.

Ordinal Data. A unidimensional quantitative continuum underlies ordinal scaling operations. Unlike in nominal scaling, the assignment of numbers reflects degree rather than kind. With the knowledge that some subjects (or units of interest) have more or less of a particular trait than others, subjects can be ordered from highest to lowest on the trait in question and numbers can be assigned consecutively to reflect this ordering. The assigned numbers, called *rank values*, are no longer just labels of convenience but instead denote a relative position in an ordered array. Collectively, rank values are data, specifically ordinal data.

It should be understood that the analytical techniques presented in this book are not generally applicable to ordinal data—at least not to ordinal data in their pristine form. There is, however, an exception: If subjects are first categorized largely on the basis of qualitative attributes (i.e., nominal scaling), and then the