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.

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

FIGURES

TABLES

1

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 (ANOVA). 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