AN INTRODUCTION TO GENERALIZED LINEAR MODELS

George H. Dunteman
Moon-Ho R. Ho

Series: Quantitative Applications in the Social Sciences

145

SAGE Publications

Series/Number 07–145

AN INTRODUCTION TO GENERALIZED LINEAR MODELS

GEORGE H. DUNTEMAN

MOON-HO R. HO
Department of Psychology, McGill University, Montreal, Quebec, Canada

Division of Psychology, Nanyang Technological University, Singapore

SAGE PUBLICATIONS
International Educational and Professional Publisher
Thousand Oaks London New Delhi

For information:

Sage Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

Sage Publications Ltd.
1 Oliver's Yard
55 City Road
London EC1Y 1SP
United Kingdom

Sage Publications India Pvt. Ltd.
B-42, Panchsheel Enclave
Post Box 4109
New Delhi 110 017 India

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Dunteman, George H. (George Henry), 1935–2004
An introduction to generalized linear models / George H. Dunteman, Moon-Ho R. Ho.
p. cm.—(Quantitative applications in the social sciences; 145)
Includes bibliographical references and index.
ISBN 0–7619–2084–6 (pbk.)
1. Regression analysis—Mathematical models. 2. Linear models (Statistics) I. Ho, Moon-Ho R. II. Title. III. Sage university papers series. Quantitative applications in the social sciences; 145.
HA31.3.D86 2006
519.5′36—dc22 2005012705

This book is printed on acid-free paper.

05 06 07 08 09 10 9 8 7 6 5 4 3 2 1

Acquisitions Editor: Lisa Cuevas Shaw
Editorial Assistant: Karen Gia Wong
Production Editor: Melanie Birdsall
Copy Editor: Daniel Hays
Typesetter: C&M Digitals (P) Ltd
Proofreader: Chloe Kristy
Indexer: Naomi Linzer

Series: Quantitative Applications in the Social Sciences

Rosarie, George E., Elizabeth, and Alyssa
—G. D.

To my parents for support, patience, and endurance
—M. H.

SERIES EDITOR'S INTRODUCTION

The course of editing this book has taken an unusual path: A change in authorship as well as editorship took place. My predecessor, Michael Lewis-Beck, was wise in seeing the value of adding to the series an introductory title on the generalized linear model. He saw through the editing of the prospectus and earlier drafts of the manuscript before stepping down as editor in early 2004. Sadly, George H. Dunteman passed away right after completing what he thought was a final draft. Further revisions were completed by Moon-Ho R. Ho, who gallantly took up the challenge and brought the manuscript to fruition with important additions and revisions to the original draft.

The outcome variables that social scientists analyze can be continuous or discrete. In our series, we have many titles that deal with the type of models represented by the classical linear regression that requires a continuous dependent variable (and a number of crucial assumptions). When the dependent variable is noncontinuous, often the probability of event occurrence is the object of a statistical model, but it can also be frequency or log frequency. During the past two decades, various forms of logit and probit (and log linear) models have become a standard issue in the social scientist's methods repertoire and the topic of quite a few titles in the series.

The relation between the two types of models—those for continuous outcome variables and those for discrete dependent variables—becomes transparent in the framework of the generalized linear model. In the social sciences, researchers are familiar and comfortable with linear or linearizable independent variables on the right-hand side of the equation, expressed as a linear combination of x and β. The dependent variable y on the left-hand side in the two types of models, however, may take on various forms, including metric, binary, ordinary, multinomial, and count. The random outcome of y in the two types of models may be distributed according to the normal, the binomial, the Poisson, the gamma distributions, among others, and all these distributions belong to the exponential family of distributions. Once we have made the proper assumption of the random distribution in y following the exponential form, the remaining task is to specify the link between the expectation of the random variable y and linear combination of x and β. This mapping of the expected random outcome variable y to the linear combination of x and β is part and parcel of the generalized linear model.

So far, we have two titles specifically discussing the generalized linear model: Gill's *Generalized Linear Models: A Unified Approach* (No. 134) and Liao's *Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models* (No. 101). The former presents the generalized linear model systematically and slightly more theoretically, and the latter provides a unified method for interpretation of estimation results from generalized linear models. The current book, however, has a more humble but nevertheless more down-to-earth goal: For the rank-and-file social science researchers who have mastered classical linear regression, how do they move from the linear regression model to the other type of models for noncontinuous dependent variables without losing sight of the common roots and similarities of the two types of models? The authors walk the reader through such process and enlighten the uninitiated about generalized linear models along the way, thus providing a good addition to the series.

—*Tim Futing Liao*
Series Editor

ACKNOWLEDGMENTS

Sage and the authors thank the reviewers of this text for their invaluable contribution.

CONTENTS

LIST OF FIGURES AND TABLES

Figures

Tables

AN INTRODUCTION TO GENERALIZED LINEAR MODELS

George H. Dunteman

Moon-Ho R. Ho
Department of Psychology, McGill University, Montreal, Quebec, Canada

Division of Psychology, Nanyang Technological University, Singapore

1. GENERALIZED LINEAR MODELS

Generalized linear models, as the name implies, are generalizations of the classical linear regression model. The classical linear regression model assumes that the dependent variable is a linear function of a set of independent variables, and that the dependent variable is continuous and normally distributed with constant variance. The independent variables can be continuous, categorical, or a combination of both. Multiple regression, analysis of variance, and analysis of covariance are examples of classical linear models. They can all be written in the form $y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon$, where y is the continuous dependent variable, X_j's are the independent variables, and ε is assumed to be a normally distributed error. The dependent variable y is decomposed into two components, a systematic component $\beta_0 + \sum_{j=1}^{p} \beta_j X_j$ and an error component ε. The systematic component is the expected value of y, $E(y)$, for a given set of values for the X_j's. The expected value of y, $E(y)$, is the mean of y, μ_y, for a given set of values for the independent variables, the X_j's; that is, $E(y|X_1, \ldots, X_p) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$. It is a conditional mean that depends on the values of the X_j's. The goal of regression analysis is to find a set of independent variables that have high explanatory power as measured through goodness of fit. This means that we can explain a large part of the variation in y by a linear combination of the independent variables. If the regression parameters, the β_j's, are large, then as the values of the X_j's change from observation to observation, the expected value of y or the conditional mean of y will vary considerably. If this variation in the conditional mean or predicted value is large relative to the variation in ε, then we have a

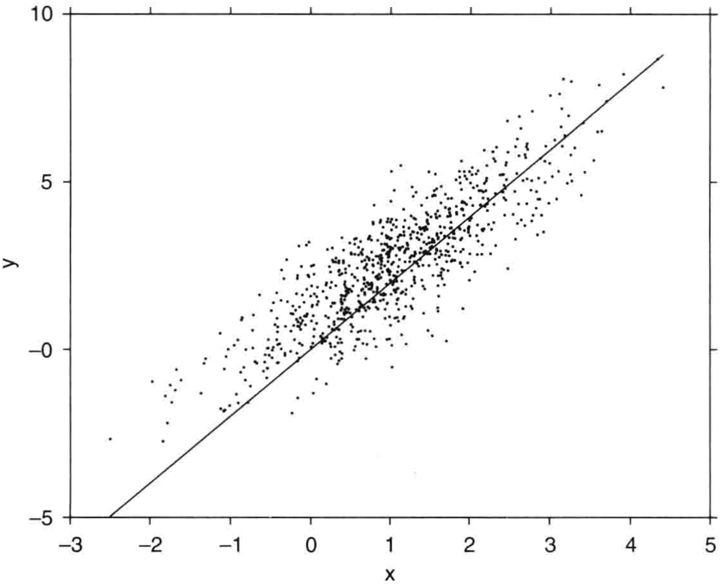

Figure 1.1 Linear Regression Model

useful model for predicting future values of y for given values of the independent variables and for understanding the relative importance of the different independent variables in explaining the variation in the dependent variable y. Figure 1.1 shows a simple linear regression model (with $\beta_0 = 1$ and $\beta_1 = 1.5$). We estimate the regression parameters, β_j's, by collecting measurements of y, X_1, X_2, ..., X_p on a random sample of observational units. For our purposes, the observational units are usually people, but in other applications the units could be anything, such as trees, cows, or even rivers. If we index the people by i and the variables by j, then we can estimate the β_j's by minimizing the error sum of squares

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 .$$

Here, subscript i is added to emphasize the fact that the values of the independent variables vary from subject to subject. This method of regression parameter estimation is commonly known as ordinary least squares.

This linear regression model has served the social sciences as well as the other sciences extremely well since its initial development in the 19th century. It is easily formulated, easy to understand, and the regression coefficients are easily estimated by ordinary least squares. Because of these factors, it is still in wide use today across all the sciences. Although it assumes normally distributed errors, it is robust when the errors are only approximately normally distributed.

Nevertheless, it has become increasingly recognized during the past several years that the linear regression model has limitations. It assumes that the dependent variable is continuous or at least quasi-continuous, such as achievement test scores, measures of personality traits, and so on. It also assumes that the continuous variable is at least approximately normally distributed and that its variance is not a function of its mean. Generalized linear models were introduced by Nelder and Wedderburn (1972) to address those limitations. Generalized linear models are a family of models developed for regression models with nonnormal dependent variables.

In many applications, the dependent variable is categorical or consists of counts or is continuous but nonnormal. An example of a categorical dependent variable is a binary variable that takes on only two discrete values, 0 or 1, where 1 indicates the occurrence of an event (e.g., dropping out of college) and 0 the nonoccurrence of an event (e.g., not dropping out of college). The goal is to model the probability of the occurrence of the event of interest. It will be shown later that logistic regression, a type of generalized linear model, is appropriate for this type of data.

An example of a dependent variable involving counts is the number of drug abuse treatment episodes in a 5-year period for a population of substance abusers. Again, it will be shown that Poisson regression, another type of generalized linear model, is appropriate for this situation. In both these cases, the dependent variable is not continuous and is far from being normally distributed. Also, 0-1 binary and count variables are nonnegative, whereas continuous dependent variables in regular regression can take on both positive and negative values.

An example of a nonnormal continuous distribution that has many applications is the gamma distribution. The gamma distribution is skewed, takes on only positive values, and its variance is a function of its mean. It is used to model a wide variety of dependent variables that can take on only positive values, such as income, survival time, and amount of rainfall. Models with gamma distributed dependent variables can be modeled within a generalized linear model framework.

It should be noted that the independent variables can take on a wide variety of distributional forms for a given distribution on the dependent variable, and they are not limited to the same distribution as the dependent variable. For example, the independent variables associated with a normally distributed dependent variable can exhibit a wide variety of nonnormal distributions, such as uniform or multimodal. As mentioned previously, regular regression assumes that whereas the mean of y varies with the independent variables, the variation of ε about the conditional means remains constant. For binary variables and count variables, the variation about the conditional mean is a function of the mean. For binary variables, the conditional mean of the dependent variables is a probability (p) (e.g., the probability of the occurrence of 1, the event), and the variation of the 0's and 1's about this mean is $p(1-p)$, which is a function of the mean (p). Because p, the mean, varies as a function of the independent variables, so does the variance of the binary variable. For count variables, the Poisson distribution is frequently used, and for this distribution the variance is equal to its mean. Therefore, as the conditional mean of the Poisson distribution varies as a function of the independent variables, so does its variance. Generalized linear models, in this case the logistic and Poisson regression models, explicitly incorporate the relationship of the mean and variance through their probability distributions in the formulation of the model and the estimation of its regression parameters.

Classical regression also assumes that the model is linear in the regression parameters. That is, it is assumed that the expected value or conditional mean is a linear function of the regression parameters. For example, $E(y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ or even $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2$. Note that the second model is linear in the parameters but nonlinear in the independent variables. In fact, classical linear regression is a specific case of a generalized linear model in which the conditional mean of the dependent variable is modeled directly rather than some transformation of the conditional mean. For other generalized linear models, the conditional mean cannot be written as a linear function of the regression parameters, but some nonlinear function of the conditional mean can be written as a linear function of the parameters; hence the name generalized linear models.

A simple example of a generalized linear model is the Poisson regression model (Figure 1.2). All the characteristics of a generalized linear model can be easily seen in this case. Moreover, it is easy to see the contrasts between this generalized linear model and a classical linear model.

In the case of Poisson regression, the expected value or conditional mean of the Poisson distributed dependent variable is

$$\lambda_i = e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}}.$$

Here, λ_i is the conditional mean of the Poisson distribution for an individual i. It is conditional in that the mean depends on the regression parameters, the β_j's, which are constant, and the specific values of the X_j's, which vary over the units of analysis (e.g., the individuals). We can compute the conditional mean λ_i for individual i by substituting his or her values of the independent variables, the X_{ij}'s, where X_{ij} is the value of the jth independent variable for individual i. In order to do this, we must have estimated the regression parameters

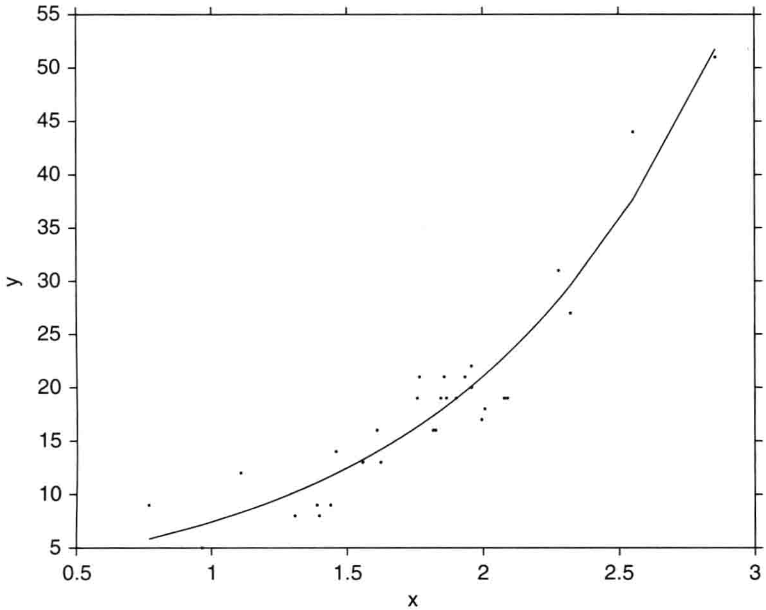

Figure 1.2 Poisson Regression Model

(the β_j's), which are unknown constants. How this is done is discussed later. We need to use the maximum likelihood method instead of least squares.

When the distribution of the dependent variable is nonnormal and its variance is a function of its mean, least squares estimates are no longer equal to maximum likelihood estimates as they are for the normal distribution. In these cases, the likelihood function must be expressed in terms of the appropriate probability density to obtain both proper parameter estimates and their standard errors. Using least squares would result in both erroneous parameter estimates and their standard errors.

The main point is that the conditional mean is not a linear function of the β_j's. If we take the natural logarithm of both sides of the Poisson regression model above, we obtain $\log_e(\lambda_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$. We have linearized the relationship between the Poisson distributed dependent variable and the independent variables by performing a nonlinear transformation on the conditional mean, λ (i.e., $\log_e(\lambda)$). We shall see that $\log_e(\lambda)$ is called the canonical link function for the Poisson regression model. It transforms the conditional mean λ of the dependent variable such that the transformed value, $\log_e(\lambda)$, is a linear function of the regression parameters. It is called canonical because $\log_e(\lambda)$ is the natural parameter of the Poisson distribution when it is expressed in exponential form. We shall also see later that the variance of a Poisson variable is equal to its mean so that if the conditional mean of the Poisson distribution increases, then so does the conditional variance associated with the conditional mean.

There are several good books on generalized linear models (Fahrmeir & Tutz, 1994; Le, 1998; McCullagh & Nelder, 1989; McCulloch, & Searle, 2001), but they usually assume a relatively high level of statistical sophistication on the part of the reader. This book assumes only basic knowledge of statistical inference and some familiarity with multiple regression. Knowledge of elementary calculus and elementary matrix algebra is not assumed, although they may be helpful in a few sections of the book. Those with little or no background in these subjects may skip or skim over those sections with little or no loss of continuity. This book is written in an informal manner and discusses the relevant statistical concepts in an intuitive manner. Its goals are to inform the reader about different types of data and allow him or her to choose the appropriate statistical model for analyzing the data and interpreting the results. In the appendix, we provide examples of how to use statistical software, SAS (SAS Institute, 2002), to fit the generalized models discussed in this book.

2. SOME BASIC MODELING CONCEPTS

We discuss the fundamental concepts of statistical modeling in the context of regular multiple regression analysis. It assumes a continuous distribution for the dependent variable with constant variance for each observation. It also assumes that the predicted value of y, its conditional mean, is a linear function of the regression parameters. We will see later that the regular multiple regression model is one of a number of specific generalized linear models if we assume that the error is normally distributed.

For three independent variables, the model can be written as $y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$, where i identifies the observation that in most applications is a person. It is assumed that ε_i has mean 0 and constant variance σ^2. In addition, it is assumed that ε_i is uncorrelated with the independent variables. The systematic component of the model is $\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$ and is the expected value of y_i or the conditional mean on the dependent variable for the ith observation given the values of X_{i1}, X_{i2}, X_{i3} for the ith observation. We express this as

$$E(y_i | X_{i1}, X_{i2}, X_{i3}) = \mu_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}.$$

The random component of the model is ε_i. We can see that as the independent variables vary, the conditional mean, μ_i, varies. The associated regression