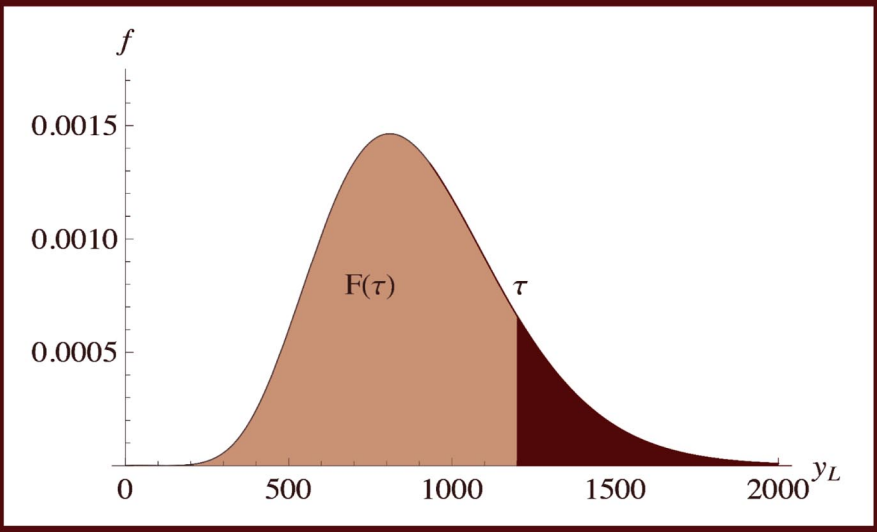
Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series

Generalized Linear Models for Categorical and Continuous Limited Dependent Variables



Michael Smithson
Edgar C. Merkle

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Generalized Linear Models for Categorical and Continuous Limited Dependent Variables

Chapman & Hall/CRC

Statistics in the Social and Behavioral Sciences Series

Aims and scope

Large and complex datasets are becoming prevalent in the social and behavioral sciences and statistical methods are crucial for the analysis and interpretation of such data. This series aims to capture new developments in statistical methodology with particular relevance to applications in the social and behavioral sciences. It seeks to promote appropriate use of statistical, econometric and psychometric methods in these applied sciences by publishing a broad range of reference works, textbooks and handbooks.

The scope of the series is wide, including applications of statistical methodology in sociology, psychology, economics, education, marketing research, political science, criminology, public policy, demography, survey methodology and official statistics. The titles included in the series are designed to appeal to applied statisticians, as well as students, researchers and practitioners from the above disciplines. The inclusion of real examples and case studies is therefore essential.

Published Titles

Analysis of Multivariate Social Science Data, Second Edition
David J. Bartholomew, Fiona Steele, Irini Moustaki, and Jane I. Galbraith

Applied Survey Data Analysis
Steven G. Heeringa, Brady T. West, and Patricia A. Berglund

Bayesian Methods: A Social and Behavioral Sciences Approach, Second Edition
Jeff Gill

Foundations of Factor Analysis, Second Edition
Stanley A. Mulaik

Incomplete Categorical Data Design: Non-Randomized Response Techniques for Sensitive Questions in Surveys
Guo-Liang Tian and Man-Lai Tang

Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists
Herbert Hoijtink

Latent Markov Models for Longitudinal Data
Francesco Bartolucci, Alessio Farcomeni, and Fulvia Pennoni

Linear Causal Modeling with Structural Equations
Stanley A. Mulaik

Multiple Correspondence Analysis and Related Methods
Michael Greenacre and Jorg Blasius

Multivariable Modeling and Multivariate Analysis for the Behavioral Sciences
Brian S. Everitt

Statistical Test Theory for the Behavioral Sciences
Dato N. M. de Gruijter and Leo J. Th. van der Kamp

Generalized Linear Models for Categorical and Continuous Limited Dependent Variables
Michael Smithson and Edgar C. Merkle

This page intentionally left blank

Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series

Generalized Linear Models for Categorical and Continuous Limited Dependent Variables

Michael Smithson
The Australian National University, Canberra, Australia

Edgar C. Merkle
University of Missouri, Columbia, Missouri, USA

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20130614

International Standard Book Number-13: 978-1-4665-5175-6 (eBook - PDF)

Dedication

Michael Smithson: To the memory of my father, James Edward Smithson (1921–1961). I think he would have liked this book.

Ed Merkle: To Lila and Grant.

This page intentionally left blank

Contents

Preface

This book is devoted to the "other" kinds of dependent variables than those for which linear regression is appropriate. These include binary, polytomous nominal, categorical ordinal, counted, interval-valued, bounded continuous, censored, and truncated variables. We argue that these dependent variables are, if anything, more common throughout the human sciences than the kind that suit linear regression. Readers acquainted with the literature on such variables will have noticed the similarity between the titles of this book and of the pioneering textbook by Long (1997). Long's book was eagerly acquired by the first author when it came out and proved an excellent source and guide over the years for both students and colleagues. Our book updates his book on topics they have in common, primarily regarding advances in special cases or extensions of models, estimation methods, model diagnostics, and of course software. Although the past two decades have seen many excellent books published on these topics, most of them are devoted to one or another specific subset of the topics. Ours is a broader but unified coverage in which we attempt to integrate the concepts and ideas shared across models and types of data, especially regarding conceptual links between discrete and continuous limited dependent variables.

At several points we bring together material that heretofore has been scattered across the literature in journal articles, book chapters, conference proceedings, software package documentation files, and blogs. Topics in our book not covered in Long's include bounded continuous variables, a greater variety of boundary-inflated models, and methods for modeling heteroscedasticity. All of the dependent variables we consider have boundaries of some kind, be they due to categorical distinctions or bounds on a continuum. The distinctions among different kinds of bounds and how to incorporate them into statistical models are fairly challenging issues, and not much guidance is available in the literature. For example, we have observed that researchers can become confused about whether boundary observations on a variable should be regarded as accurate scores or censored values. Throughout the book we guide the reader to appropriate models on the basis of whether the bounds are inherent in a construct or variable, or imposed (e.g., by censoring or truncation). Likewise, although both the concepts and software are available for dealing with heteroscedasticity, it remains a relatively neglected topic in the applied statistical literature despite its considerable importance. Heteroscedasticity is especially relevant for the kinds of dependent variables we deal with here, both because it can frequently arise in the data and because some models for these variables are inherently heteroscedastic. We therefore treat both kinds of heteroscedasticity: Unconditional in the sense that it is

due to the bounded nature of the construct or variable, and conditional on values or states of independent variables.

Wherever possible, we have illustrated concepts, models, and techniques with real or realistic datasets and demonstrations in R (R Development Core Team, 2013) and Stata. Each substantive chapter also has several exercises at the end. Both illustrations and exercises are intended to help readers to build conceptual understanding and fluency in using these techniques.

Data and Software

We illustrate the models and methods in this book using both R (R Development Core Team, 2013) and Stata software. We elected to use these pieces of software through a combination of personal preference, popularity, and access. The data files used in this book are all freely available; to obtain them, R users can install the `smdata` package that is freely available on CRAN. The installation, followed by loading the package and loading datasets, can be completed with the following R commands.

```
## Install package
install.packages("smdata")

## Load package
library("smdata")

## Load, e.g., the email dataset within smdata
data("email")
```

On the topic of R, we also note that Thompson (2009) has written a valuable manual that describes the use of R for categorical data analysis; it can be freely obtained at `https://home.comcast.net/~lthompson221/Splusdiscrete2.pdf`.

Stata users and others can obtain the data from `https://dl.dropbox.com/u/1857674/SmiMerBook/SmiMerDoc.html`, which includes the data in both Stata format and csv format. The page also includes details about each data file and some extra code. Finally, Stata users will find the book on categorical and limited dependent variables by Long and Freese (2006) very useful.

Acknowledgments

We are indebted to numerous people who have helped to improve this book. We received substantial and very useful feedback on chapter drafts, illustrative examples, and supplementary material from Yiyun Shou, Jay Verkuilen, four anonymous reviewers, and the Australian National University psychology Honours and graduate students participating in statistical workshops during 2012. We also received useful feedback and discussion from graduate students in Ed's GLM courses at Wichita State University and from colleagues in the department of Psychological Sciences at the University of Missouri. Roger Koenker and Achim Zeileis provided prompt re-

sponses to our questions regarding the nuances of various packages and code, and the maintainers of R and CRAN aided us with the generous contribution of their time. Likewise, Bill Rising at Stata, as part of their author support program, made valuable suggestions for improving our attempts at Stata code. On the data side, Lukas Hulsey provided the medical treatment preference data that were used in Chapter 4, Justin Owens provided the eye-tracking data that were used in Chapter 5, Michael Gurr's Honours thesis data was used in Chapter 6, Ken Mavor contributed data used in Chapter 7, and Stephen Tang contributed data used in Chapter 8. At Chapman and Hall, John Kimmel gave us expert editorial advice, Shashi Kumar provided essential LaTeX support, project coordinators Kathryn Everett and Kate Gallo guided us throughout the book's gestation, and project editor Marsha Hecht provided invaluable production assistance. Without them this book would not have been possible, let alone presentable. Of course, any errors or flaws remaining in this book are our responsibility.

The book was typeset by the authors in LaTeX using the `chapmono` class. We also made extensive use of the `apacite` package for citations and indexes. We are indebted to the developers and maintainers of this free software.

Finally, we each owe ineradicable debts to our families, most especially to our wives, Susan and Victoria. Susan, yet again, tolerated the late nights, working weekends, obsessional preoccupation, and spousal neglect that unfortunately accompanies my book-writing (and she knows I can't even promise that this is the last one!). Victoria also tolerated these things, independently of book-writing to some degree. She additionally had a baby during the middle of book-writing, making for a very busy time. She continued to keep good spirits and flutter her eyelashes during this busy time, however, and I am grateful.

This page intentionally left blank

List of Figures

List of Tables

Notation

Symbol	Definition
$\sim$	is distributed as (e.g., $X \sim N(0,1)$)
$\approx$	is approximately equal to (e.g., $\pi \approx 3.14$)
α_j	the intercept/cutpoint associated with the j^{th} logit of an ordinal regression model
$\text{Be}(\omega, \upsilon)$	a Beta distribution with parameters ω and υ
$\text{B}(\cdot)$	Beta function
$\boldsymbol{\beta}$	a vector of regression coefficients
$\boldsymbol{\beta}_j$	a vector of regression coefficients associated with the j^{th} logit of an ordinal regression
β_k	the regression coefficient associated with x_k
χ^2_{df}	a chi-squared-distributed variate with df degrees of freedom
$D(M)$	The deviance associated with Model M
$\exp(x)$ or e^x	The exponential of x
$f(\cdot)$	usually a probability density function (pdf)
$F(\cdot)$	usually a cumulative density function (cdf)
$g(\cdot)$	usually a link function in the location submodel of a GLM
G^2	Likelihood ratio statistic
$\Gamma(\cdot)$	Gamma function
$h(\cdot)$	usually a link function in the dispersion submodel of a GLM
I	number of unique combinations of values that may be assumed by a vector of predictors $\boldsymbol{x}$
J	number of categories in a categorical or ordinal variable
$L(\boldsymbol{\theta}\|\boldsymbol{y},\boldsymbol{X})$	A model likelihood function
$\ell(\boldsymbol{\theta}\|y_i,\boldsymbol{x}_i)$	Individual i's contribution to the likelihood
μ	population mean
$\text{N}(\mu,\sigma)$	a normal distribution with mean μ and standard deviation σ
n	number of observations
φ	population precision in Beta and Negative Binomial Models
$\phi(\cdot)$, $\Phi(\cdot)$	the pdf and cdf for the standard normal distribution $N(0,1)$
π	mixture probability in zero-inflated models
$P(\cdot)$	Probability of an event or a proposition
p	Usually a probability in the role of a dependent variable in a GLM
q	the number of estimated model parameters
r	a model residual

σ	population standard deviation
$\boldsymbol{\Sigma}$	a population variance–covariance matrix
t	exposure variable
τ	censoring or truncation threshold in probit, logit, and Tobit models
$\boldsymbol{\theta}$	a model parameter vector
ω, υ	parameters of a beta distribution
X^2	Pearson goodness-of-fit statistic
$\boldsymbol{X}$	a matrix of predictor variables
$\boldsymbol{x}_i$	a vector of predictor variables for observation i
x_k	the k^{th} predictor variable
y	a response variable
y^*	a latent, continuous variable underlying y
W	Wald statistic

About the Authors

Michael Smithson is a Professor in the Research School of Psychology at The Australian National University in Canberra. He received his PhD from the University of Oregon. His primary research interests are in judgment and decision making under uncertainty, statistical methods for the social sciences, and applications of fuzzy set theory to the social sciences. He has authored or co-authored six books and co-edited two volumes. His other publications include more than 140 refereed journal articles and book chapters.

Edgar C. Merkle is an Assistant Professor in the Department of Psychological Sciences at the University of Missouri. He received a PhD in Quantitative Psychology and an MS in Statistics, both from the Ohio State University. His research interests include latent variable models, subjective probabilities and forecasts, and statistical computing. He has authored numerous journal articles within these areas.

This page intentionally left blank

Chapter 1

Introduction and Overview

1.1 The Nature of Limited Dependent Variables

Many variables in the social sciences are "limited" in the sense that their supports have boundaries. In fact, we claim that the vast majority of these variables are bounded. These bounds consist of two related kinds: Categorical boundaries and bounds on a continuum. The key distinction between these is that cases contained within categorical bounds are treated as identical in value or state, whereas cases falling in a bounded continuum may take different values within that range.

The primary rationale for this book and related books, including Long (1997), Agresti (2012), Powers and Xie (2008), and Bishop, Fienberg, and Holland (1975), is that the most popular distributions for model error (the normal and t distributions) assume that the dependent variables to which they are applied are unbounded (i.e., their support encompasses the entire real line). In contrast, bounded data require distributions that take their bounds into account.

Categorical bounds imply discreteness, and so categorical random variables require discrete distributions. These occupy the first part of this book. Bounds on continuous variables raise important measurement issues, primarily regarding cases at or near the boundaries. Are such cases accurately recorded or do their true scores lie "beyond the bounds?" Often the bounds are artifacts insofar as they are imposed by such considerations as constraints on the number of items comprising a scale or a test, a practical need to identify scale endpoints, or a decision to record only cases that exceed or lie below some threshold.

We distinguish three situations regarding bounds. First, the dependent variable data are completely known for all cases, so that cases on a boundary have been accurately measured and, in this sense, the boundary is "real." Second, some of the cases are *censored* because they are only partially known (e.g., we know that a particular debtor owes less than $1,000 but we do not know exactly how much she owes). Third, some cases are *truncated* because they have been excluded from the sample on the basis of some characteristic (e.g., a bank records only losses exceeding $1,000). We will examine censoring and truncation in Chapters 5 and 7.

Chapter 6 deals with doubly bounded continuous dependent variables, i.e., those that have both a lower and an upper bound, where the boundary cases are real. The most obvious examples are proportions and percentages, but other examples are readily found such as borrowings on credit cards with upper limits, rating scales used by

judges in performance sports such as diving or gymnastics, prototypicality or degree of membership rating scales, and happiness or satisfaction ratings.

Bounds may be "unreal" in two ways. First, they may be imposed arbitrarily, such as the endpoints in the popular agree–disagree rating scales. If the endpoints are "strongly disagree" and "strongly agree" then all we know about cases occupying those endpoints is that they are that strong or stronger in their degree of (dis)agreement. Another influence on the veridicality of scale endpoints is the granularity of the scale. A confidence rating scale will have more accurate endpoints if it uses 20 bins than if it uses 4.

Second, the bounds may be real (e.g., a score of 0% or 100% on an exam) but the boundary cases may or may not be censored depending on the measurement purpose. A score of 0% or 100% on an exam is a true score if the exam includes all relevant test items, but not if it is considered to include only a sample of relevant items. Likewise, if a survey asking respondents to report the amount of time they spent eating during the last 24 hours turns up cases reporting 0, those cases may be taken as true scores if the purpose is simply to measure how much time people from the relevant population spent on eating during the past 24 hours, but not if it is to measure the proportion of 24 hours that people typically devote to eating.

Finally, limited dependent variables may be "boundary inflated" in the sense that the proportion of boundary cases exceeds that presupposed by the distribution model. A count variable such as number of cigarettes smoked in the past month might produce a large number of zeros in a population where most people do not smoke cigarettes. Chapters 5 and 6 include treatments of boundary-inflated models.

1.2 Overview of GLMs

Generalized linear models (GLMs), as originally described by Nelder and Wedderburn (1972) and expanded upon by McCullagh and Nelder (1989), form the foundation of this book. In this section, we define GLMs, relate them to simpler models, and discuss extensions.

1.2.1 Definition

GLMs are extensions of the standard linear regression model to situations where the dependent variable (or response variable) is limited. A linear predictor is common to all models, but the way in which the linear predictor relates to the data is different. To begin, consider the standard linear regression model, which we will generally call the *Gaussian GLM* for reasons that will become clear:

$$y = \boldsymbol{x}\boldsymbol{\beta} + e = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K + e, \tag{1.1}$$

where we assume a 1 in the first entry of $\boldsymbol{x}$ and that $e \sim \mathrm{N}(0, \sigma^2)$. With $K+2$ or more observations, we can fit this model to the data and obtain estimates of $\boldsymbol{\beta}$ and σ^2. Given estimates of $\boldsymbol{\beta}$ and some value of $\boldsymbol{x}$, we can then predict the mean of y. The regression weights $\boldsymbol{\beta}$ provide insight about the extent to which the predicted mean of y is impacted by each of the x_j. Thus, we typically wish to interpret the βs and test

whether they differ from zero. These tests and interpretations help us to summarize the general impact of the x_j on y.

To "generalize" the model in Equation (1.1), we must think about it in an alternative manner. We consider three features of the model: (i) the distribution associated with y, (ii) the parameter of the distribution on which we wish to focus, and (iii) the way in which we model the parameter via a linear predictor. For the Gaussian GLM, we have (i) a normal distribution for y, (ii) focus on μ, the normal distribution's mean parameter, and (iii) a linear predictor placed directly on μ. The third feature is admittedly confusing for this model, because it is not clear why it is needed. However, it will help us soon.

In the spirit of the three features that we just noted, we could rewrite the model from (1.1) as

$$y|\boldsymbol{x},\boldsymbol{\beta} \sim \mathrm{N}(\mu,\sigma^2) \tag{1.2}$$

$$\mu = \boldsymbol{x}\boldsymbol{\beta}. \tag{1.3}$$

These equations no longer contain an error term that is added to a linear predictor. Instead, we assume a distribution for y and then place a linear predictor on the conditional mean of the distribution of the response variable. The distribution on y, the normal (also known as the Gaussian), leads us to call this model a Gaussian GLM. The parameter that we model, μ, is unbounded and can therefore be directly modeled via the linear predictor. In general, however, model parameters are not unbounded, so that a linear predictor may make nonsensical predictions. For example, we will see many situations where a model parameter can only assume values between 0 and 1. In this case, placing a linear predictor directly on the parameter may result in predictions that are less than zero or greater than 1. We need a function that "unbounds" the model parameter, making it sensible to use a linear predictor. This function is termed a *link function*, because it "links" a model parameter to a linear predictor. Given a distribution for y, there exists a special link function that implies some good statistical properties (involving the fact that the sufficient statistics for the model are of a simple form). This special link function is termed a *canonical link function*. However, choice of link function is more often guided by precedent than by whether or not it is canonical.

We will describe a variety of link functions throughout this book, though we illustrate only one here. In Chapter 2 and elsewhere, we model a probability parameter p that can only assume values in $(0,1)$. To use a linear predictor in this model, the link function must transform values in $(0,1)$ to values in $(-\infty,\infty)$. A common choice (which also happens to be the canonical choice) is the logit link, also known as the log-odds link. This is given as

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right). \tag{1.4}$$

This function is described in much more detail in Chapter 2. In a GLM context, we would place a linear predictor on $\mathrm{logit}(p)$ to obtain

$$\mathrm{logit}(p) = \boldsymbol{x}\boldsymbol{\beta}, \tag{1.5}$$

so that the regression weights lead to predictions on the transformed, logit scale. We typically wish to have predictions on the probability scale, and we can obtain these via an inverse link function. The inverse link transforms unbounded predictions back to the bounded scale, with the inverse logit given as

$$\text{logit}^{-1}(\boldsymbol{x\beta}) = \frac{1}{1+\exp(-\boldsymbol{x\beta})}. \tag{1.6}$$

The main point that the reader should take from this discussion is the fact that, for many distributions, we must unbound a parameter before using a linear predictor. This is accomplished via a link function. Throughout this book, we typically describe GLMs via the three features above: the assumed distribution for y, the focal model parameter (which is usually the conditional mean of the distribution), and the link function that associates the focal parameter with a linear predictor. The near-exclusive focus on linear predictors largely follows the tradition of model developments; however, at least for unbounded data, a linear predictor generally approximates a function of interest via a first-order Taylor-series expansion (see, e.g., Venables, 2000). The accuracy of this approximation obviously varies.

1.2.2 Extensions

As defined by McCullagh and Nelder (1989), GLMs include only models whose distributions arise from the exponential family (see, e.g., Casella & Berger, 2002 for a formal definition of the exponential family). While this family includes many common distributions (including the normal, binomial, Poisson, beta, and gamma), many of the models that we describe in this book do not fall in the family. This makes the title of the book inaccurate in some respects, although the models not included in the exponential family still use many of the same concepts. We describe here some extensions of "plain" GLMs, some of which still fall in the GLM family and some of which do not. While we recognize the inaccuracy in terminology, we generally ignore the "GLM family or not" distinction because it does not have a major impact on the applied researcher.

We will often distinguish between *location* and *dispersion* parameters throughout the book. Informally, location parameters are those that influence the central tendency of a distribution. Dispersion parameters, on the other hand, influence only the variability of a distribution. When a parameter influences both central tendency and variability, it is generally regarded as a location parameter (the mean of the Poisson distribution is one example).

Applied researchers are quite familiar with modeling location: most popular statistical tests employ null hypotheses associated with mean parameters. Further, many of the models in this book will allow us to assess the impact of predictor variables on the mean of a distribution. Dispersion parameters, on the other hand, are often regarded as error parameters or nuisance parameters, and less attention is typically devoted to them. As we will see in this book, however, these parameters are often very important both for estimation and interpretation. Dispersion parameters can affect the interpretations that we make about location parameters, and they can also

be interesting to study in isolation. Expanding on the latter issue, there are some situations in which we would expect the predictor variables to directly influence the variability of a response variable. For bounded variables, this variability can additionally be interpreted as polarization. The associated statistical tests of polarization can inform many theories in the social sciences. Additionally, the mean and variance of a bounded variable are not independent of one another. Thus, dispersion parameters generally play an important role in models for bounded dependent variables.

For both location and dispersion parameters, there exist a few common scenarios where the attributes of y are not a good match with the associated model. In these scenarios, there are often relatively simple modifications that can be employed to salvage the original model. We describe the location scenarios separately from dispersion scenarios below.

Location. Focusing on location parameters, the scenarios primarily include situations where y is *censored* or *truncated*. Truncation occurs when we exclude some values of y above or below a certain point. For example, if we ask current smokers about the number of cigarettes that they smoke each month, we have excluded responses of "zero." This situation may be called "truncation from below at 1" or, alternatively, "left truncation at 1." In contrast, we would observe truncation from "above" (or from the "right") if we excluded values above a certain point of the response variable y.

Censoring is similar to truncation, except that all values of y above or below a point τ have simply been relabeled as τ. For example, imagine a bathroom scale with a maximum weight of 400 lbs. While this will not be an issue for many people, anyone who weighs more than 400 lbs will be assigned a weight of 400. This is a case of "right-censoring at 400." Censored observations are observed (e.g., a 450-lb individual is recorded as weighing 400 lbs), while truncated observations are unobserved (e.g., people who smoke zero cigarettes are not included in the sample).

Censoring and truncation are often observed in the context of count regression (Chapter 5), though we also encounter these issues in other situations. Chapter 7 describes the general handling of censoring and truncation, including situations where the associated threshold is unknown (for example, where we do not know the maximum weight that a scale can record).

Dispersion. Focusing on dispersion parameters, major scenarios are *overdispersion* and *heteroscedasticity*. When we have overdispersion, the model-predicted variance systematically underestimates the variance observed in y. That is, the model simply cannot account for the variance observed in y while simultaneously accounting for the mean (i.e., for the location) of y. The overdispersion issue is important because, when it occurs, the standard errors associated with the location estimates are too small. This can often lead the researcher to erroneously infer statistically significant effects of a predictor variable on y (i.e., Type I errors). To correct for overdispersion, it is possible to add an extra parameter that accounts for the variance that the original model could not. It is also sometimes possible to use a different distribution for y that better handles dispersion. The overdispersion issue is most relevant to the count