CHAPTER 1. INTRODUCTION AND OVERVIEW

1.1 Overview of This Book

This book provides a course on generalized linear models for bounded variables. We focus on numeric dependent variables whose scales are bounded either at one end or both ends. Examples are income (typically bounded below at 0), hours spent on an activity per day (bounded between 0 and 24), or percentage of a population eligible to vote (bounded from 0 to 100).

Why is this topic important? The human sciences deal in many variables that are bounded. Ignoring bounds can result in misestimation and improper statistical inference. On the other hand, taking bounds into account not only can provide more accurate statistics but also often reveals insights that otherwise would escape the researcher's notice.

Why is a book like this needed? Bounded scales and variables often are analyzed with conventional but inappropriate techniques. This is due to lack of familiarity with the techniques for working with bounded variables. Books and other teaching materials that introduce generalized linear models ignore or give only scant attention to such techniques. Instead, techniques and relevant concepts regarding bounded variables typically crop up only in specialized contexts such as survival analysis, psychophysics, engineering, or econometrics. Our goal is to make them generally available and accessible to nonspecialists.

There are book-length surveys of limited dependent variables, including the now classic monograph by Long (1997b) and the more recent treatment by Smithson and Merkle (2013). Books also have been devoted to subsets of these variables. Categorical variables understandably have received the most attention, including handbooks on logistic regression (Hosmer Jr., Lemeshow, & Sturdivant, 2013), log-linear analysis (Agresti, 2013), and ordinal categorical regression (Agresti, 2010). Singly bounded variables usually appear under the guise of "life distributions" (Marshall & Olkin, 2007). Censoring and truncation have been given an accessible introduction by Breen (1996). To our awareness, however, there is no introduction to quantitative bounded variables that is both accessible and as thorough as might be desired. By focusing exclusively on these variables (rather than including them along with categorical variables), we have attempted such an introduction in this book.

This is a good time for this book to appear, because the early part of the 21st century has seen rapid growth in the availability of software resources for analyzing bounded variables. Bounded-variable models are available in popular computing environments, including R, Stata, MPlus, and SAS. Generalized linear models (GLMs) for singly bounded variables have been widely available for some time in these environments. GLMs for doubly bounded variables have been made available in R, Stata, and SAS. The stage is set for accelerating understanding and application of these models, and this book is intended to provide a guide and impetus for such applications.

This is a "second" course in generalized linear models. We assume that the reader is familiar with linear multiple regression and has had at least an introduction to the general linear model. The Sage collection has excellent books on both topics: Colin and Michael Lewis-Beck's (2015) *Applied Regression* and Jeff Gill's (2000) *Generalized Linear Models*.

Here is what our book covers. In this first chapter, we introduce the different types of bounds and present examples of two of the most well-known kinds. In Chapter 2, we focus on singly bounded variables, beginning with the basic concepts underpinning a GLM for such variables and then introducing readers to three of the most popular relevant distributions and examples of models that use them. In Chapter 2, we also discuss alternative methods for dealing with cases on the boundary and how to choose such a method. In Chapters 3 and 4, we cover methods for dealing with doubly bounded variables. Chapter 3 focuses on models employing the beta distribution, while Chapter 4 introduces distributions for modeling quantiles of doubly bounded variables. In Chapter 5, we deal with what are known as "censored" and "truncated" variables, starting with the popular Tobit model and then moving on to different types of censoring and truncation, as well as non-Gaussian and heteroscedastic models. In Chapter 6, we conclude the book with an overview of bounded variables and brief discussions of extensions to the models covered in the preceding chapters. These extensions include multivariate and random-effects models, Bayesian estimation, and the treatment of bounded covariates or independent variables.

We provide worked examples in every chapter, using real data sets from a variety of disciplines. The software used for the examples include R, SAS, and Stata. The data, software code, and detailed explanations of the example models are available in the supplementary materials on the website for this book: www.study.sagepub.com/researchmethods/qass/smithson-and-shou-generalized-linear-models. We take care not to include particulars in the book about software that quickly become outdated, instead relegating such specifics to the supplementary web-based materials where they can be updated as needed. The supplementary materials also contain additional details, including code, development, and interpretation of the models, in some cases going beyond the models described in the book itself.

1.2 The Nature of Bounds on Variables

Bounds on variables occur in two contexts: as categorical bounds and bounds on one or more continuous ranges. This book deals with variables that have bounds on continuous ranges. Moreover, we restrict our treatment to variables that have only one or two bounds. Bounded variables also require two kinds of considerations regarding GLMs for them. First, there is the problem of constructing a GLM that takes the bounds into account (e.g., by not generating predictions outside of the range). Second, there is the issue of how to model and interpret cases at the bounds.

A useful typology of bounds on continuous variables distinguishes among three kinds: "absolute," "censoring," and "truncating." Absolute bounds are values beyond which it is impossible for the variable to go (e.g., a proportion cannot go below 0 or above 1). Censoring bounds are values that only put a lower or an upper limit on the true scores of cases on the boundary (i.e., those cases' scores are "censored"). For example, a webpage automatically logs visitors out after their inactivity has exceeded 15 minutes, so lengths of visits to that webpage are only known to be at least 15 minutes' duration if they timed out. Truncating bounds are those for which cases are excluded altogether from a sample (e.g., a perception experiment excludes participants whose vision is worse than 20-40). Absolute bounds more strongly constrain the choice of distributions for modeling the data because they determine the support or domain of the distribution, whereas censoring and truncation do not entirely determine the distribution support.

Among the most common kinds of censoring or truncating are when bounds on a variable are an artifact of a nonexhaustive collection of items whose scores are combined to form a scale or due to scale endpoints that truncate scores. An example of the first kind is the ethical risk-taking subscale in the DOSPERT (the Domain-Specific Risk-Taking Inventory; Weber, Blais, & Betz, 2002), composed of eight items, each of which has a score from 1 to 5, so that the subscale range is from 8 to 40. We cannot infer that a person who scores 8 on this scale never takes any ethical risks, because the eight items do not exhaust the list of unethical acts. An example of the second is a scale recording household incomes with an upper bound "more than $*I*," where *I* is a threshold amount.

There also are different varieties of truncation due to sample selection. Boundary cases may consist exclusively of exclusions from a sample, such as amounts owing on household mortgages in which the zeros are those households that are not currently mortgaged. Conversely, boundary cases may include a mix of sample exclusions and inclusions. For instance, data consisting of the number of cigarettes smoked by a person in the past week may include zeros that are smokers as well as zeros that are nonsmokers.

The question of whether bounds are absolute or not can be debatable for at least three reasons. First, the answer may depend on the target construct being measured by the scale. Percentage score on an examination is a commonplace example. Clearly, it makes no sense for a percentage to fall below 0 or above 100 in this context, and if the exam is intended simply to measure knowledge about what is being examined, then these bounds are absolute. However, if the examination items do not exhaust the subject of the examination, then the bounds are not absolute in the sense of measuring the student's knowledge of the subject. A student scoring 0 still may have some subject knowledge, and a student scoring 100 may not know everything about the subject.

Second, the definition of the construct may determine the nature of the bounds. In Tobin's (1958) original application of a censored regression model, the dependent variable was the ratio of household expenditure on durable goods to disposable income. If we regard the underlying construct as propensity to purchase durable goods, then we may elect to define an observed value of 0 as corresponding to a latent propensity of 0 (so the boundary is absolute) or to define it as a censored latent propensity.

Third, the boundaries may be open to multiple interpretations by respondents. An example is a scale with verbal anchors, such as a World Values Survey (World Values Survey Association, 2015) item that asks participants, "How important is it for you to live in a country that is governed democratically?" on a scale from 1 to 10, where 1 is labeled *not at all important* and 10 is labeled *absolutely important*. Taken literally, this scale's bounds seem absolute insofar as values outside these bounds do not make sense. However, there is no guarantee that all respondents will interpret these verbal anchors literally.

Finally, we return briefly to the issue of boundary cases. A preponderance of cases on a boundary produces what is sometimes called a "boundary-inflated" distribution. It suggests that at least some of the boundary cases may be distinct in some way from the rest of the cases. In some contexts, there are good reasons for supposing the existence of such a distinction. Suppose we ask people to estimate the probability that humanity will become extinct within the next thousand years. A large number of zeros in the responses would indicate that there may be two types of respondents, those who believe the human species never will become extinct and those who believe that extinction is possible. We may then decide to ascertain what distinguishes the zero-respondents from the others, separately from identifying predictors of how probable people think human extinction is. Another important class of boundary cases is so-called corner solutions, as identified in the economics literature. A corner solution occurs when an agent maximizes his or her utility at a boundary, as in a person who refuses all bets, no matter how attractive, on grounds that it is against his or her moral code to gamble. We will revisit the treatment of boundary cases several times in this book. First, however, we will briefly review the concepts from the generalized linear model that will be used throughout this book.

1.3 The Generalized Linear Model

1.3.1 Definitions and Concepts

Many of the models described in this book are generalized linear models, or GLMs (McCullagh & Nelder, 1989). We are going to provide a brief introduction to them, starting with definitions and basic concepts. For a complete introduction to GLMs, however, readers should consult Jeff Gill's book (2000). First, let us consider the *general linear model*. Suppose we have a dependent variable, Y, whose expected value is a linear function of predictor variables x_j, for $j = 1, \ldots, J$:

$$Y = \mu + e = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_J x_J + e = \boldsymbol{x\beta} + \epsilon, \quad (1.1)$$

where the vector $\boldsymbol{x}$ has $x_0 = 1$ as its first entry, $\boldsymbol{\beta}$ is the vector of coefficients, and ϵ has a normal distribution with mean 0 and variance σ^2. Another way to write the model in equation (1.1) is to think of Y as having a *conditional distribution* (i.e., a distribution whose parameters are at least partly determined by the predictor variables). So, in equation (1.2), we describe the conditional distribution of $Y|\boldsymbol{x}, \boldsymbol{\beta}$ as normal with mean μ and variance σ^2, where μ is determined by the weighted linear combination of predictors, $\boldsymbol{x\beta}$. This last equation in equation (1.2) is the *systematic component* of the general linear model, and the ϵ error term in equation (1.1) is what makes $Y|\boldsymbol{x}, \boldsymbol{\beta}$ the *stochastic component*.

$$\begin{gathered} Y|\boldsymbol{x}, \boldsymbol{\beta} \sim N\left(\mu, \sigma^2\right) \\ \mu = \boldsymbol{x\beta}. \end{gathered} \tag{1.2}$$

The *generalized linear model* often is confused with the general linear model, but it is indeed a more general form of a linear model in two respects. First, it relaxes some of the assumptions required of the general linear model and admits other distributions than the normal. Second, it involves a *link function* connecting the systematic and stochastic components of the general linear model. The link function, g, is applied to the parameter μ being estimated via the weighted linear combination of predictors, as shown in equation (1.3). This function is smooth and monotonic in μ. In some representations of the GLM, the inverse is used instead (i.e., $\mu = g^{-1}(\boldsymbol{x\beta})$) because it focuses on the estimation of the parameter μ. The general linear model, then, is a special case of the GLM, with the link function being the identity: $g(\mu) = \mu$.

$$\begin{gathered} Y|\boldsymbol{x}, \boldsymbol{\beta} \sim f\left(\mu, \sigma^2\right) \\ g\left(\mu\right) = \boldsymbol{x\beta}. \end{gathered} \tag{1.3}$$

In some cases, there may be more than one available link function. A special link function is known as the *canonical link*, because it arises as a result of how the distribution f is defined. Why is this distinction between canonical and noncanonical link functions important? Gill (2000) provides a detailed and accessible explanation of the desirable statistical properties of the canonical link. Aside from the statistical or theoretical reasons, the most pragmatic reason is that in a few cases, the domain of the canonical link function is not the same as the permissible range of μ. For example, this is true of the exponential and gamma distributions, whose canonical link functions are the reciprocal functions. In such cases, modelers often will use an appropriate noncanonical link function (such as the log) that restricts estimates to the permissible range.

As an example, we introduce the lognormal GLM, where the log of Y is assumed to have a normal distribution. We are using this example because the lognormal is a popular choice for modeling variables that have a lower bound of 0, such as income or reaction time. In equation (1.4), we see that Y is distributed as lognormal (LN) with parameters

μ and σ, and the link function $g(\mu) = \log(\mu)$. Although μ can only take positive values, $\log(\mu)$ can take any value on the real line, positive or negative. Thus, g unbounds the μ parameter so that it can be estimated by the weighted linear combination of predictors without worrying about the lower bound.

$$\begin{gathered} Y|\boldsymbol{x}, \boldsymbol{\beta} \sim LN\left(\mu, \sigma^2\right) \\ \log\left(\mu\right) = \boldsymbol{x\beta}. \end{gathered} \tag{1.4}$$

It is important to bear in mind that the link function does not always translate straightforwardly to the expected value of Y or its variance. In the lognormal distribution, although $E(\log(Y)) = \mu$, the expectation of Y is $E(Y) = \exp(\mu + \sigma^2/2)$. Likewise, although the variance of $\log(Y)$ is σ^2, the variance of Y is $(\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)$. This type of difference between the link function and the parameterization of summary statistics such as the mean, variance, and quantiles will crop up frequently for bounded variables, and we will return to this issue several times to extract the connection between it and the nature of bounds on variables.

Most of the GLMs in this book involve *location-scale* distributions that are fully specified by two parameters. Often, these consist of a location parameter (e.g., a mean) and a scale or dispersion parameter (e.g., a standard deviation). Roughly speaking, a location parameter determines the central tendency of a distribution, whereas a dispersion parameter determines the variability of a distribution.

Traditionally, applied modelers have focused exclusively on modeling the location parameter, relegating the dispersion parameter to the status of a nuisance parameter or a role in evaluating location model error. This tradition stems from two related sources. One is the popular assumption of homoscedasticity or homogeneity of variance (i.e., variance is constant regardless of the location of the mean), as in conventional linear regression models.

The second, more implicit source is the fact that many location-scale distributions whose support is the real line (e.g., the normal and t distributions) have the property that the location parameter can be changed without altering dispersion and vice versa. However, for variables with bounds, this generally is not the case. When location and dispersion are not independent of one another, then modeling both location and dispersion parameters becomes important.

Returning to the lognormal GLM example, we already have seen that in its original scale, the mean and variance of Y are influenced both by μ and by σ. Thus, in its original scale, Y is naturally heteroscedastic.

It therefore could make sense to model both μ and σ explicitly, with predictor variables for each. We then would have two *submodels*, as shown in equation (1.5), a *location submodel* for μ and a *dispersion submodel* for σ, each with its own link function (in this example, it is the log for both). Moreover, in $\boldsymbol{z\delta}$, the variables in the vector $\boldsymbol{z}$ may differ from or overlap with those in $\boldsymbol{x}$.

$$\begin{aligned} Y|\boldsymbol{x}, \boldsymbol{\beta} &\sim LN\left(\mu, \sigma^2\right) \\ \log(\mu) &= \boldsymbol{x\beta} \\ \log(\sigma) &= \boldsymbol{z\delta}. \end{aligned} \tag{1.5}$$

1.3.2 Estimation

Most of the GLMs in this book are estimated via maximum likelihood (ML) estimation rather than least squares estimation. The latter has been the default method in linear regression, and in linear regression, least squares and ML estimates are identical. However, in many GLMs, especially those that model dispersion parameters, ML estimation is relatively straightforward whereas least squares estimation often is inapplicable. Equation (1.1) presents a “least squares” view of a GLM with the conventional error term ϵ, whose sum of squares is minimized in least squares estimation. On the other hand, equations (1.2) to (1.5) present a “likelihood” view that refers directly to a distribution, conditional on its parameter values. The error term ϵ has been absorbed into the conditional distribution, and by the time we arrive at equation (1.5) with its submodel for σ, there is no viable place for an error term of that kind.

Instead, the ML approach deals with the likelihood of each observation in the distribution, conditional on the parameter estimates. Suppose we have a random sample of independent observations x_i, for $i = 1, \ldots, N$, from a random variable X whose probability density function is $f(x, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of parameters that define the density function. Then the likelihood of any x_i conditional on μ and σ is $f\left(x_i \,|\boldsymbol{\theta}\right)$. The *likelihood function* is the product of all of these likelihoods, due to the independence of the x_i. The log of the likelihood function is more computationally convenient to work with, so the ML approach conventionally uses the log-likelihood function

$$L\left(\boldsymbol{x}\,|\boldsymbol{\theta}\right) = \sum_{i=1}^{N} \log\left(f\left(x_i \,|\boldsymbol{\theta}\right)\right). \tag{1.6}$$

The *maximum likelihood estimates* of the parameters in a GLM are the values of those parameters that maximize $L\left(\boldsymbol{x}\,|\boldsymbol{\theta}\right)$.

1.3.3 Evaluating and Comparing Models

Given alternative models with different sets of parameters and predictors, researchers frequently wish to compare these models to determine which of them fit the data best. This inclination naturally arises from the availability of likelihood statistics and some of their attractive properties, and while it may not always result in the best practice regarding comparisons among models, we will use it throughout this book because it is practical and also related to Bayesian model comparison methods such as Bayes factors.

To begin, it is important to distinguish between *nested* and *nonnested* models. Model 1 is nested in Model 2 if Model 2 includes all of Model 1's parameters and additional parameters as well. For instance, suppose that Model 1 has a location submodel $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, while Model 2's location submodel is $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_2 x_3$ and Model 3's location submodel is $\mu = \beta_0 + \beta_2 x_2 + \beta_2 x_3$. Then Model 1 is nested in Model 2 but not in Model 3, whereas both Models 1 and 3 are nested in Model 2. Sometimes the nested model (e.g., Model 1) is called the *reduced model*, and the nesting model (Model 2) is called the *full model*.

When models are nested, they may be compared via likelihood-based tests. There are three asymptotically equivalent such tests: the likelihood ratio, Lagrange multiplier, and Wald tests (Engle, 1984). The likelihood ratio test (LRT) may be written as

$$G_{12}^2 = -2\left(L\left(\boldsymbol{y}\,|\boldsymbol{x}_1, \boldsymbol{\theta}_1\right) - L\left(\boldsymbol{y}\,|\boldsymbol{x}_2, \boldsymbol{\theta}_2\right)\right), \tag{1.7}$$

where the subscripts 1 and 2 denote the parameters in Models 1 and 2, respectively, and all of the variables and parameters in Model 1 are contained in those for Model 2. When the null hypothesis that these model likelihoods are equal (i.e., the additional parameters in Model 2 do not improve model fit) is true, G_{12}^2 asymptotically follows a χ^2 distribution, whose degrees of freedom are $k_2 - k_1$, where k_1 and k_2 are the number of parameters in Models 1 and 2, respectively. Rejecting the null hypothesis motivates the researcher to regard Model 2 as better fitting than Model 1. Failing to reject the null hypothesis, on the other hand, justifies preferring Model 1 over Model 2 on grounds of parsimony.

The other two tests sometimes are useful alternatives to the LRT, partly because each of them requires full estimation of only one of the alternative models. The Lagrange multiplier test requires only that the reduced model be estimated. The full model is then fitted in a single iteration, and the resulting change in fit is used to evaluate the full model.

The Wald test, on the other hand, requires only complete estimation of the full model. It simultaneously tests null hypotheses regarding parameters in the model, enabling the researcher to discard those parameters for which the null hypothesis cannot be rejected. However, the Wald test uses two approximations (the χ^2 approximation and approximate standard errors for the parameters), whereas the LRT uses only the χ^2 approximation, so the Wald approach is more susceptible to small-sample inaccuracy.

These three tests cannot be applied to comparisons between nonnested models. The two most popular statistics for comparing nonnested models are the *Akaike information criterion* (AIC; Akaike, 1974) and the *Bayesian information criterion* (BIC; Schwarz, 1978), sometimes also known as the Schwarz information criterion. The AIC is defined as

$$\mathrm{AIC} = -2L\left(\boldsymbol{y}\,|\boldsymbol{x},\boldsymbol{\theta}\right) + 2k, \tag{1.8}$$

where k is the number of parameters estimated in a model, and the BIC is defined as

$$\mathrm{BIC} = -2L\left(\boldsymbol{y}\,|\boldsymbol{x},\boldsymbol{\theta}\right) + k\log\left(N\right), \tag{1.9}$$

where N is the sample size. Both the AIC and BIC penalize a model by its number of parameters (i.e., its complexity). The lower the information criterion score, the better the model fit is relative to the model's complexity. The BIC penalizes model complexity more severely than the AIC by including sample size as a factor, and some researchers prefer the BIC to the AIC because they believe the BIC is more likely to help them avoid selecting an overfitted model. On the other hand, while the AIC actually is based on information theory, the BIC is not. Vrieze (2012) presents a thoughtful discussion of alternative information criteria.

In addition to computing overall measures of fit with the data, researchers are well advised to evaluate the model in greater detail. Broadly speaking, there are three ways that this can be done: investigating the extent to which a candidate model fits all of the observations equally well, ascertaining whether particular observations have a disproportionate influence on the model parameter estimates, and assessing how sensitive the model fit is to perturbations of the parameter estimate values. These are related but distinct investigations. Assessments of how well a model fits individual observations typically are conducted using residuals, which are based on the difference between an observation's value and the model's prediction thereof. Assessments of the influence an observation has on the model parameter estimates usually are done via “leverage” statistics, which measure the sensitivity of parameter estimates to the presence or perturbation of the observation. Assessments of model fit sensitivity to perturbations of model parameter values are less common than the other two types of assessment but can be done as a by-product of the other two kinds or simply by computing likelihoods of the model with alternative parameter values sampled within their respective standard error or confidence bounds.

Turning first to residuals, four commonly employed kinds are the raw (or response), Pearson, Anscombe, and deviance residuals. The response residuals are just the difference between the observations and the model’s predictions:

$$r_i = y_i - y'_i \left| \left(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}} \right) \right. . \tag{1.10}$$

The *Pearson residuals* scale the raw residuals by the standard error of the predicted value:

$$r_{pi} = \frac{r_i}{\sqrt{V\left(y'_i\left(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}\right)\right)}}, \tag{1.11}$$

where V is the variance function associated with the GLM’s distribution. Pearson residuals are asymptotically normally distributed under appropriate conditions, but in real situations, they can be strongly skewed. *Anscombe residuals* (Anscombe, 1953) transform the numerator and denominator of the Pearson residuals to make them unimodal and symmetric, thereby as close to a normal distribution as possible. A full explanation of these residuals would constitute a digression here, but Gill (2000) provides an accessible and detailed explanation.

A different type of residual is the deviance residual. The *deviance* statistic associated with a model is defined as twice the difference between the log-likelihood of a “saturated” model that perfectly predicts the observations y_i and the log-likelihood of the model:

$$D\left(\boldsymbol{y}\,|\boldsymbol{x}, \boldsymbol{\theta}\right) = 2\left(L\left(\boldsymbol{y}\,|\boldsymbol{\xi}\right) - L\left(\boldsymbol{y}\,|\boldsymbol{x}, \boldsymbol{\theta}\right)\right), \tag{1.12}$$

where $\boldsymbol{\xi}$ is the vector of parameters that perfectly predicts the y_i. The *deviance residual*, then, is

$$r_{di} = \mathrm{sign}\left(r_i\right)\sqrt{d_i\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right)}, \tag{1.13}$$

where sign() takes the sign of the term in the parentheses, and d_i denotes the ith contribution to the deviance,

$$d_i\left(y_i\,|\boldsymbol{x}_i, \boldsymbol{\theta}\right) = 2\left(L\left(y_i\,|\xi_i\right) - L\left(y_i\,|\boldsymbol{x}_i, \boldsymbol{\theta}\right)\right). \tag{1.14}$$

In addition to examining residuals, assessing the effect of individual observations on the model provides a second way of evaluating the model. A commonplace assumption among researchers is that excessively influential observations also are outliers, but this is not always true. Thus, there is a role for *influence statistics*, which measure the impact that an observation has on the model. The main idea underpinning influence statistics is cross-validation, whereby one observation at a time is removed from the data and the model is fitted to the remaining data. The most popular influence statistic is Cook's distance, which measures the sum of the changes in regression coefficients when one observation is removed from the data (Cook, 1977). Cook's distance was derived in the context of linear regression, and variants of it have subsequently been developed for other GLMs. Likewise, a popular standardized measure of change in a model parameter when an observation is removed is the change in the coefficient divided by the standard error of the original parameter estimate, for example,

$$\Delta_i\left(\hat{\beta}\right) = \left(\hat{\beta} - \hat{\beta}_{(i)}\right)\Big/\hat{\sigma}_{\hat{\beta}}, \tag{1.15}$$

where $\hat{\beta}_{(i)}$ denotes the ML estimate of β when the ith observation has been removed. These influence statistics are called "dfbetas."

1.4 Examples

Having reviewed GLMs and introduced some ideas about bounds, we will finish our introduction with two examples. We hope these examples will make these concepts more concrete and also illustrate some of the benefits from using techniques that take bounds into account.

1.4.1 Absolute Bounds Example

Income distributions are a prototypical example of a distribution with one bound. We will use household income data for households with positive income for two of the years, 2010 and 2015, from the American Community Survey database (U.S. Census Bureau, 2015), for our example. These data comprise 148,076 households, and their income distribution and normal quantile-quantile (Q-Q) plot are shown in Figure 1.1. The distribution is strongly skewed (the largest household income exceeds $2 million), and the Q-Q plot shows that it deviates far from a normal distribution.

Figure 1.1 Household Income Distribution and Q-Q Plot

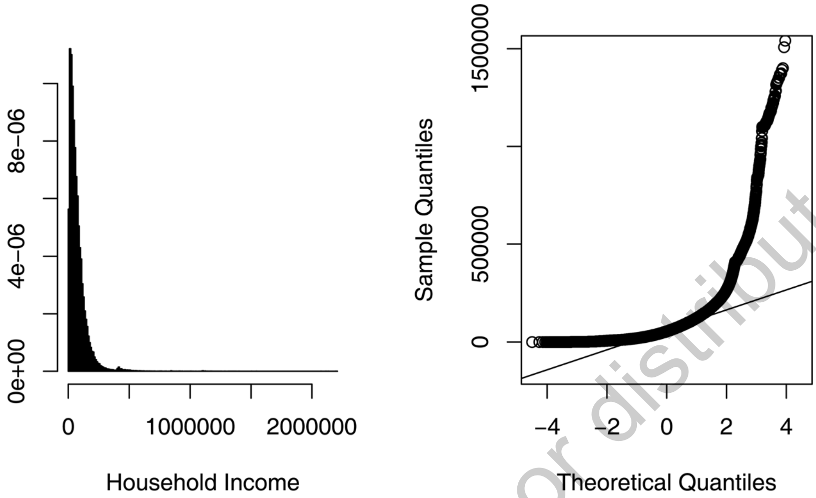
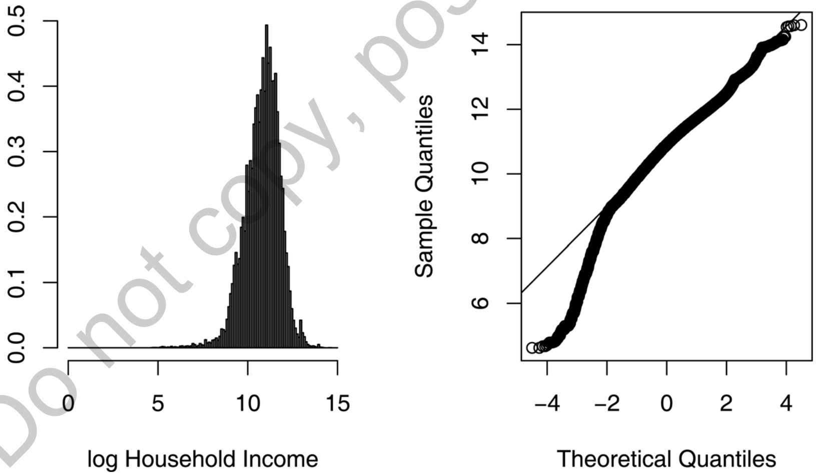


Figure 1.2 Log(Household Income) Distribution and Q-Q Plot



The distribution of the log income and its Q-Q plot in Figure 1.2 suggest that the lognormal may be an appropriate distribution for modeling these data. The Q-Q plot shows that the log income distribution closely corresponds to the normal distribution for all but approximately 2% of the data in the lower tail (i.e., the data from about 2 standard deviations below the mean downward).

Suppose we wish to compare the incomes for households that obtained food stamps with those that did not, and we also would like to ascertain whether the difference between them changed from 2010 to 2015. A linear regression model (equivalent to an analysis of variance [ANOVA] with Type III sums of squares) yields this outcome:

$$\begin{gathered} Y|\boldsymbol{x}, \boldsymbol{\beta} \sim N\left(\mu, \sigma^2\right) \\ \mu = \boldsymbol{x\beta} = 75990.63 - 54389.33x_1 + 10029.23x_2 - 4476.95x_1x_2, \end{gathered} \tag{1.16}$$

where x_1 is the food-stamp dummy variable ($x_1 = 1$ for households obtaining food stamps and 0 for households with no food stamps), x_2 is the year dummy variable ($x_2 = 1$ for 2015 and 0 for 2010), and x_1x_2 is the product of x_1 and x_2.

Now, there are two commonplace versions of a "lognormal" model among researchers. One of these is a linear regression model with the log of the dependent variable. Let us denote this as the "log-DV" model. This model with the same predictors yields this outcome:

$$\begin{gathered} \log(Y)|\boldsymbol{x}, \boldsymbol{\beta} \sim N\left(\mu, \sigma^2\right) \\ \mu = \boldsymbol{x\beta} = 10.888 - 1.223x_1 + 0.101x_2 + 0.053x_1x_2 \\ \log(\sigma) = \delta = -0.101. \end{gathered} \tag{1.17}$$

All of the coefficients are statistically significant, not least because of the enormous sample size. Our focus here is on the structure, effects, and goodness of fit of each model. Beginning with the structure, we can see from the coefficients that there is one obvious difference between the two models, namely that the linear model has a negative interaction term and the log-DV model has a positive interaction term. It also should be borne in mind that the magnitudes of the coefficients differ considerably due to the fact that one model is in the linear scale and the other is in the log scale. There are several more or less equivalent ways to compare the effects of these two models, all of which require transforming from one model's scale to the other's. We will transform the log-DV model's estimates to the linear scale.

Table 1.1 displays the predicted means for each of the models in their respective scales. In the row below the log-DV model's means, the corresponding means have been computed in the linear scale. Recall that although $E(\log(Y)) = \mu$, the expectation of Y is $E(Y) = \exp(\mu + \sigma^2/2)$.

Table 1.1 Linear and Log-DV Model-Predicted Means and Effects

	Food Stamp	No	No	Yes	Yes
	Year	2010	2015	2010	2015
Linear model					
$\hat{\sigma} = 78,270$	Means	75,990.63	86,019.85	21,601.30	27,153.58
			Ratios	3.52	3.17
Log-DV model					
$\hat{\sigma} = 0.9036$	Means	10.8878	10.9883	9.6652	9.8188
	Linear scale	80,505.23	89,012.98	23,705.75	27,642.44
			Ratios	3.40	3.22

For example, the 2010 no-food-stamps mean is $E(Y) = \exp(10.8878 + 0.9036^2/2) = 80,505.23$ (allowing for a small roundoff error). The estimated standard deviation, $\hat{\sigma}$, for each model is displayed in the left-most column. Although the log-DV model's transformed means differ from those of the linear model, the effect sizes are fairly similar. The linear model's mean of the no-food-stamp households in 2010 is 3.52 times higher than the food-stamp household mean for the same year $(75,990.63/21,601.30 = 3.52)$, and the corresponding log-DV ratio of its means is 3.40. For 2015, the linear model means ratio is 3.17 and the log-DV means ratio is 3.22.

Now we turn to goodness of fit. Both models perfectly reproduce the sample means in their respective scales. That is, the linear regression model perfectly reproduces the means in the original scale, and the log-DV GLM perfectly reproduces the means in the log scale. However, as we saw earlier, the log-DV model does not produce the same mean in the linear scale as the linear model does (e.g., the mean income for households not receiving food stamps in 2010 is 80,505.23 rather than 75,990.6).

Now let us turn to the second version of a lognormal model, in which the link function for the mean response is the log. This model yields a somewhat different set of coefficients from those in the log-DV model of equation (1.17):

$$Y|\boldsymbol{x}, \boldsymbol{\beta} \sim LN\left(\mu, \sigma^2\right)$$
$$\log(\mu) = \boldsymbol{x\beta} = 11.2384 - 1.2579x_1 + 0.1240x_2 + 0.1048x_1x_2. \tag{1.18}$$

We have deliberately chosen a "trivial" example to illustrate the fact that transforming the lognormal model's estimates back to the linear scale reproduces the sample means in the linear scale, just as the linear model does. For example, using a more precise estimate of the intercept to diminish roundoff errors, the model's estimated mean income for households not receiving food stamps in 2010 is $\exp(11.238365) = 75,990.6$. Moreover, the lognormal model has exactly the same log-likelihood as the linear model. Thus, the most important practical difference between the log-DV and lognormal models is that the log-DV model rescales the dependent variable, whereas the lognormal model rescales the model.

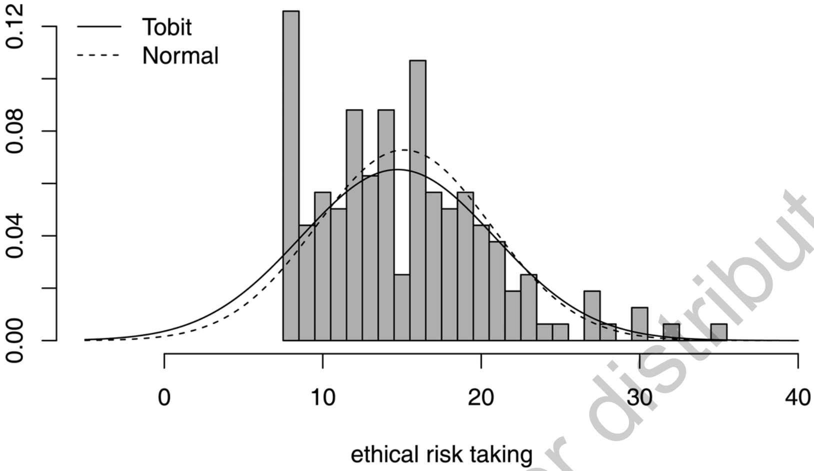
1.4.2 Censoring Bounds Example

We now turn to an example of a model for a censored dependent variable. A class of 159 second-year Australian National University psychology students completed the DOSPERT, and we will model their responses on the "ethical risk-taking" subscale. Items in this subscale ask respondents to rate their likelihood to commit acts such as cheating on an exam and pirating software. A substantial number of the students' responses yielded the lowest possible score. Why is this lowest possible score "censored"? One way to think about this is that it may be possible to find other examples of unethical acts that even some of these 20 respondents would rate themselves as likely to do. After all, the DOSPERT includes only eight examples of unethical acts, nowhere near exhausting the human repertoire.

Figure 1.3 displays a histogram of the subscale responses, showing 20 cases on the lower boundary of the subscale. Two fitted distributions also are displayed. The dashed-line distribution is a normal distribution that reproduces the sample mean, $\hat{\mu} = 15.12$, and standard deviation, $\hat{\sigma} = 5.48$. The solid-line distribution is a normal distribution fitted via a *Tobit model*, which takes censoring into account (Tobin, 1958). It estimates the mean as $\hat{\mu} = 14.73$ and standard deviation as $\hat{\sigma} = 6.11$, appropriately decreasing the mean and increasing the standard deviation estimates. Moreover, the Tobit model's distribution fits the data better than the normal distribution: The log-likelihood for the Tobit is -471.1 whereas for the normal, it is -495.6.

The Tobit model is the most popular GLM for censored outcomes. It assumes that the data are sampled from a censored normal distribution. In our example, that means assuming that the 20 boundary cases, if they were uncensored, would be distributed as in the lower tail of a normal distribution whose support extends beyond the scale's boundary (the lower tail of the solid-line distribution in Figure 1.3). Censoring can occur at either end of a scale (or both), and there are several kinds of censoring. These matters will be elaborated in Chapter 6, but here

Figure 1.3 Normal and Tobit Distributions Fitted to Subscale Scores



we focus on a lower-censored Tobit model. The traditional notation for such a model describes the uncensored observations in terms of equation (1.1), that is,

$$y_i = \boldsymbol{x_i\beta} + \epsilon_i, \tag{1.19}$$

where $\epsilon_i \sim N\,(0, \sigma)$. For the variable's censored observations, suppose τ is the censoring threshold. Then

$$\boldsymbol{x_i\beta} + \epsilon_i \leq \tau, \tag{1.20}$$

so that $\epsilon_i \leq \tau - \boldsymbol{x_i\beta}$ and therefore

$$\Pr(y_i \leq \tau\,|x_i) = 1 - \Phi((\boldsymbol{x_i\beta} - \tau)/\sigma_i), \tag{1.21}$$

where Φ is the standard normal cumulative distribution function. A model for the censoring rate is given by equation (1.21).

We have seen in our example that the Tobit model gives different estimates of the mean and standard deviation from those of the linear regression model. The linear regression model's estimate of the mean is just $\boldsymbol{\mu_x\beta}$, where $\boldsymbol{\mu_x}$ denotes the vector mean of the predictor variables. The Tobit model's mean estimate is

$$\mathrm{E}\,[y] = \Phi_\tau \boldsymbol{\mu_x\beta} + \sigma\phi_\tau + \tau(1 - \Phi_\tau), \tag{1.22}$$

where $\Phi_\tau = \Phi((\boldsymbol{\mu_x\beta} - \tau)/\sigma)$ and $\phi_\tau = \phi((\boldsymbol{\mu_x\beta} - \tau)/\sigma)$. We now consider a linear regression and a Tobit model with scores on the DOSPERT

Table 1.2 Linear Regression and Tobit Model Summaries

Model	Coefficient	Estimate	Standard Error		
Regression	Intercept	1.414	1.274	Log-likelihood	−449.4
	Health risk	0.687	0.062	*t*	11.132
Tobit	Intercept	−0.592	1.464	Log-likelihood	−424.4
	Health risk	0.770	0.070	*t*	10.968
Quantiles	10%	25%	50%	75%	90%
Empirical	8	11	14	18	22
Regression	10.35	12.41	14.48	17.22	19.42
Tobit	9.41	11.72	14.03	17.11	19.57

health subscale predicting scores on the ethical risk-taking subscale. The two models' coefficients, *t* statistics, predicted versus empirical quantiles, and log-likelihoods are displayed in Table 1.2. The linear regression model's coefficient is smaller than the Tobit's, although their *t* values are similar (due to the Tobit model's larger residual standard error estimate). As before, the log-likelihoods indicate that the Tobit model is the better fit. This also is borne out by the predicted quantiles, which in all but one case are closer to their empirical counterparts for the Tobit than for the regression model.

In this chapter, we have tried to set the stage for the rest of the book by reviewing the general linear model, introducing the concepts needed for understanding the nature of bounds on variables, and presenting what we hope are motivating demonstrations of models for bounded variables. We now move on to considering models for singly bounded variables, when the boundary is treated as absolute.