Psychological Methods
2006, Vol. 11, No. 1, 54–71

Copyright 2006 by the American Psychological Association
1082-989X/06/$12.00 DOI: 10.1037/1082-989X.11.1.54

A Better Lemon Squeezer? Maximum-Likelihood Regression With Beta-Distributed Dependent Variables

Michael Smithson
The Australian National University

Jay Verkuilen
University of Illinois at Urbana–Champaign

Uncorrectable skew and heteroscedasticity are among the "lemons" of psychological data, yet many important variables naturally exhibit these properties. For scales with a lower and upper bound, a suitable candidate for models is the beta distribution, which is very flexible and models skew quite well. The authors present maximum-likelihood regression models assuming that the dependent variable is conditionally beta distributed rather than Gaussian. The approach models both means (location) and variances (dispersion) with their own distinct sets of predictors (continuous and/or categorical), thereby modeling heteroscedasticity. The location submodel link function is the logit and thereby analogous to logistic regression, whereas the dispersion submodel is log linear. Real examples show that these models handle the independent observations case readily. The article discusses comparisons between beta regression and alternative techniques, model selection and interpretation, practical estimation, and software.

Keywords: beta distribution, regression, variance, generalized linear model, heteroscedasticity

The normal-theory regression model is, unarguably, the workhorse of applied statistics, and it is broadly applicable to many problems in practice. Simply put, normally distributed errors turn out to be a pretty good assumption in many situations. However, this is not universally true, and in some areas of research violations of normality are common. Numerous devices have been invented to render data suitable for this approach in the face of assumption violations. Even so, there are circumstances in which these remedial measures cannot help or are undesirable for substantive or theoretical reasons. Uncorrectable skew, heteroscedasticity, and multimodality of the dependent variable are among the most common difficulties. Although many psychological researchers commonly believe that normal-theory regression (or equivalent models, such as analysis of variance [ANOVA]) is robust against violations of assumptions—which indeed it is in many circumstances—this view often is mistaken and can lead to misinterpretations of data and theoretical impasses. It is particularly true given commonly used measures such as survey responses with bounded response sets or proportions.

The following example highlights the ways in which conventional regression can yield misleading results. The data were supplied by K. Pammer in the School of Psychology at The Australian National University (Pammer & Kevan, 2004). We consider the relative contribution of nonverbal IQ and dyslexic versus nondyslexic status to the distribution of 44 children's scores on a test of reading accuracy. The reading score distributions for the two groups of children are shown in Figure 1. These scores have been linearly transformed from their original scale to the open unit interval (0, 1) by first taking $y' = (y - a)/(b - a)$, where b is the highest possible score on the test and a is the smallest possible score, and then compressing the range to

Michael Smithson, School of Psychology, The Australian National University, Canberra, Australian Capital Territory, Australia; Jay Verkuilen, Department of Psychology, University of Illinois at Urbana–Champaign.

Software for performing the analyses developed in this article is available as code in SAS, S Plus/R, and SPSS on the Web at http://dx.doi.org/10.1037/1082-989X.11.1.54.supp and at http://www.anu.edu.au/psychology/people/smithson/details/Index.html.

This research was supported by National Research Service Award MH14257 from the National Institute of Mental Health to the University of Illinois. The research was conducted while Jay Verkuilen was a predoctoral trainee in the Quantitative Methods Program at the Department of Psychology, University of Illinois at Urbana–Champaign. Both authors contributed equally to this article.

We wish to acknowledge the comments of Carolyn Anderson, Winifred Louis, Ken Mavor, Craig McGarty, Carol Nickerson, and Alecia Santuzzi on earlier versions of this article and the participants at the Quantitative Division Seminar, Department of Psychology, University of Illinois at Urbana–Champaign.

Correspondence concerning this article should be addressed to Michael Smithson, School of Psychology, The Australian National University, Canberra ACT 0200, Australia. E-mail: michael.smithson@anu.edu.au

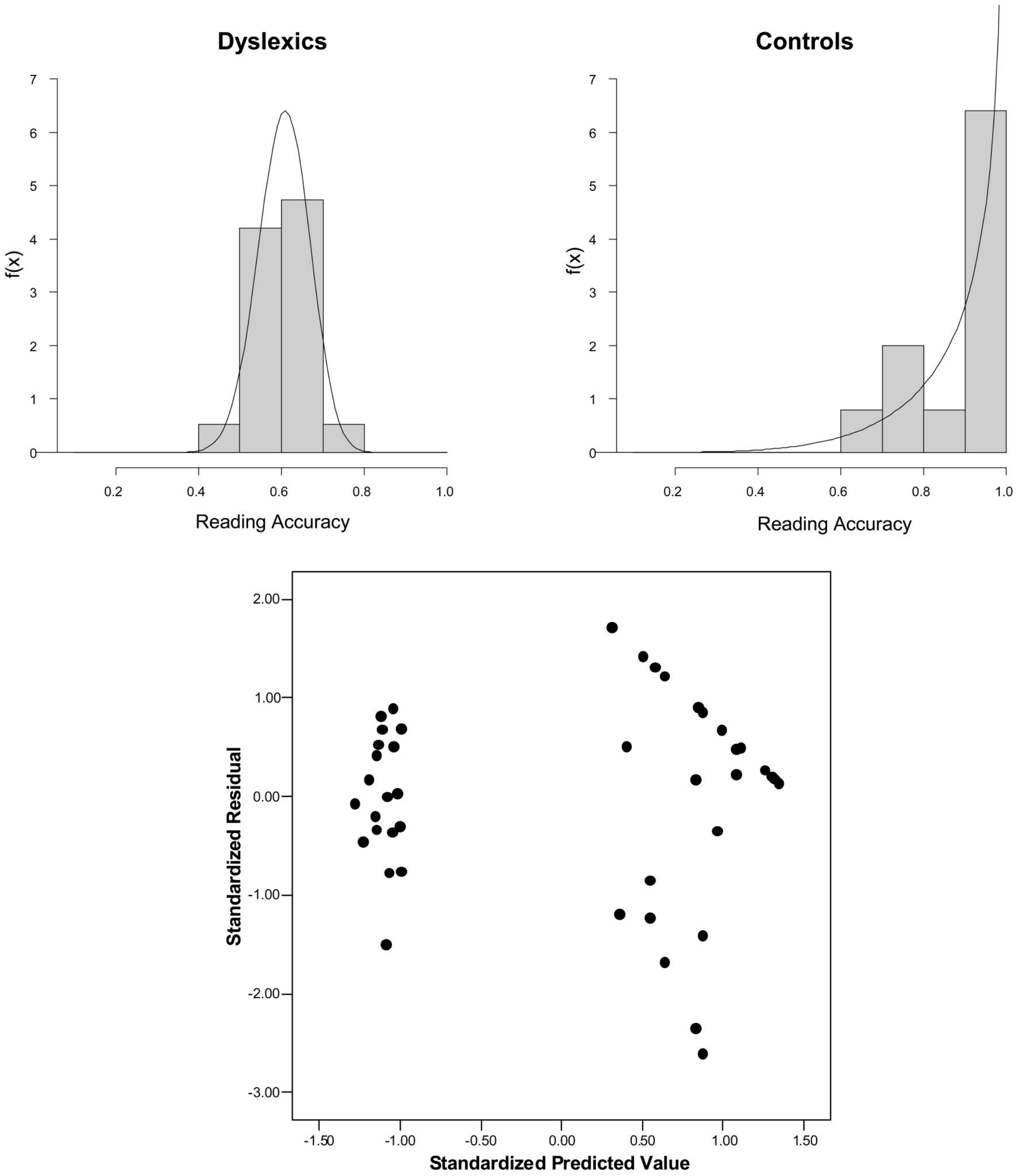


Figure 1. Reading accuracy scores for dyslexics and controls and ordinary least squares residuals.

avoid zeros and ones by taking $y'' = [y'(N - 1) + 1/2]/N$, where N is the sample size (the rationale for these rescalings is presented in the first part of the next section and is further detailed in materials referred to in Footnote 2).

The nondyslexic readers' mean accuracy score is .899, whereas the mean for readers who have dyslexia is .606. The third plot in Figure 1 shows the residuals for a linear regression model, including the group, IQ, and an interaction term for IQ by group. The plot clearly suggests there is heterogeneity of variance, and Levene's test concurs with this assessment, $F(1, 42) = 12.434$, $p = .001$. Moreover, the usual regression approach does not permit us to model this difference in dispersion between the groups. More important, even when accuracy scores are transformed using the logit, a linear regression model using standardized IQ scores, effects-coded dyslexia status, and an interaction term produces a significant main effect for dyslexia status ($p < .0001$), a nonsignificant main effect for IQ ($p = .12$), and a marginally nonsignificant interaction effect ($p = .07$). The impression is that IQ makes little or no discernible independent contribution. As will be shown in our third example, this impression is misleading because these regression models confound effects of covariates on means with effects on variances.

Skew and heteroscedasticity are also theoretically and substantively important in their own right. Although location (mean) shifts are routinely considered "the" measure of effect, it is also possible that a causal factor manifests primarily through variation and therefore heteroscedasticity. For instance, gender differences in variability on ability and achievement tests have long been noted, with males usually showing greater variability on tests of numerical and spatial abilities but no variability differences in tests tapping verbal abilities (Feingold, 1992; Maccoby & Jacklin, 1974). As Willingham and Cole (1997, pp. 51–52) observed, these differences in variability exaggerate a male advantage in the upper tail and a male disadvantage in the lower tail. They also point out that gender differences in variability may be just as important as differences in means.

Furthermore, in a bounded response scale location, variation and skew are not so neatly separated as they are in the unbounded scale of the normal distribution, so a location shift might well imply skew, for instance in the presence of an active floor or ceiling. The theoretical relevance of skewness and dispersion arises in many domains in psychology. The dyslexia example is a case in point, as are many case-control studies in clinical psychology. Typically, one or the other population is strongly skewed on the variable of interest, and the variability of scores is of both theoretical and practical interest. For example, Louis, Mavor, and Terry (2003) critiqued methodological practice in social psychology on these grounds. Among other points, they argued that variables measuring degree of adherence to social norms often are strongly skewed, and normal-theory regression models consistently underestimate the contribution of those variables to predictive models of behavior and attitudes even when "appropriately" transformed. The upshot is that norms are passed over in favor of individual differences as explanatory variables.

We present a maximum-likelihood regression technique based on the beta distribution. The beta distribution is a very flexible two-parameter family that can accommodate skew, a form of bimodality, and symmetry not unlike a normal distribution. Conditionally, that is, in a regression model, the beta distribution can accommodate even more shapes. Its main assumptions are that the dependent variable may be regarded as continuous, interval-level, and bounded between two known endpoints. Many kinds of data in the social sciences are reasonably assumed to have these properties—indeed, assuming normality carries with it continuity and interval-level measurement, so this assumption is already implicit in most analyses. However, many scales and questionnaire items are bounded in nature, whereas the domain of the normal distribution (i.e., the range of values for which its density is defined) is unbounded. The domain for the beta distribution, by contrast, is bounded, and our regression models for it naturally respect these bounds, which may be of theoretical importance. The beta regression model is also especially useful for proportions. Regression models for proportions are frequently difficult, particularly when the values are centered near one of the boundaries of the unit interval.

Though not commonly used by psychologists, there are remedial measures for heteroscedasticity in the context of the normal-theory regression, such as the Huber–White heteroscedasticity-consistent covariance estimator, commonly referred to as the *sandwich estimator*, or the many computationally intensive robust regression techniques developed over the last three decades (Long & Ervin, 2000; Stautde & Sheather, 1990; Wilcox, 2005).[1] Substantive approaches such as maximum-likelihood models for normal regression with multiplicative dispersion covariates (Greene, 2000, pp. 517–521) or the random slopes model in multilevel analysis (Snijders & Bosker, 1999) attempt to model the heteroscedasticity explicitly. Like the latter two methods, the beta regression approach presented here is more than just a remedy. It permits researchers to model dispersion explicitly and naturally, particularly for bounded variables displaying skew. Theories and hypotheses concerning dispersion and heteroscedasticity can be tested directly.

There is a small literature on beta regression, all focusing exclusively on the beta as a model for proportions. Usually the beta distribution is used for other purposes, most commonly in modeling univariate distributions for a variety of phenomena and as a conjugate prior for proportions in Bayesian models. It is instructive that in the otherwise exhaustive *Handbook of Beta Distributions,* there is no mention of beta regression (Gupta & Nadarajah, 2004) despite the fact that the beta function itself dates back to Newton and the beta distribution has been used extensively in statistics for more than a century. We also draw readers' attention to the fact that there is a sizable parallel literature on beta-binomial regression (an early instance is Crowder, 1978), suited to modeling proportions of counts, where the beta distribution is used as a hierarchically specified random effect.

The first three examples of beta regression come out of the literature on organizational economics and public management. Brehm and Gates (1993) modeled police compli-

[1] It should be noted that it is possible—and indeed often desirable—to use the sandwich estimator with nonlinear and generalized linear models, particularly when misspecification is suspected but there is no obvious remedy for it, for example, when there seems to be unobserved heterogeneity among the participants but no variable exists with which to model it. However, efficiency is lost when it is used, and it is conservatively biased in small samples, sometimes substantially so. See Hardin and Hilbe (2003), especially pp. 28–32, and references therein.

ance with supervision and used the standard parameterization of the beta distribution, which has two shape parameters. Unfortunately, the standard parameterization complicates the formulation of a regression model and makes interpretation difficult. Paolino (2001) used the same mean-dispersion parameterization that we adopt in this article, which greatly simplifies interpretation. Buckley (2002) implemented a simple Bayesian version of Paolino's model and estimated it via Markov-chain Monte-Carlo (MCMC) procedures. Apparently independent of these authors, Ferrari and Cribari-Neto (2004) derived a similar beta regression model using Fisher scoring, which recently has been implemented in the SAS GLIMMIX procedure (SAS Institute, 2005). However, they did not model the dispersion but instead treated it solely as a nuisance parameter. This is, in our view, a major oversight, as the ability to model dispersion can be shown to be very useful. Again, seemingly independently of the aforementioned work, Kieschnick and McCullogh (2003) compared the performance of a beta regression model for proportions, as used in economics and finance research, with several alternatives and concluded that it is often the best option. They also mentioned that the econometrics package SHAZAM (Version 7) contains a model using the identity link with a beta-distributed dependent variable. For reasons that will become apparent, those authors do not favor that approach and neither do we.

The beta regression is directly related to an extended generalized linear models framework for joint modeling of means and dispersions described in chapter 10 of McCullagh and Nelder (1989). Smyth (1989) developed this framework for random variables in the exponential family of distributions. The beta distribution forms a two-parameter exponential family, so the framework is easily adapted to it, and most of the machinery in beta regression models will be familiar to researchers accustomed to the generalized linear models, particularly logistic regression and log-linear models. Researchers not familiar with the generalized linear model should consult references such as Liao (1994), Long (1997), or Powers and Xie (2000) for introductions. Fahrmeir and Tutz (2001) or McCullagh and Nelder (1989) are more advanced texts providing important technical results.

The beta regression technique is readily implemented in standard statistical packages. Given a general program for maximum-likelihood estimation or nonlinear regression, programming a beta regression is straightforward. We have successfully implemented the model in SAS, Splus/R, SPSS, and Mathematica.[2] Other authors have implemented beta regression models in Gauss (Paolino, 2001), Stata (Buckley, 2002), R (Ferrari and Cribari-Neto's location-only model in a package by de Bustamante Simas in 2004), SAS (Ferrari and Cribari-Neto's [2004] model in the GLIMMIX procedure, which can also handle random effects via penalized quasi-likelihood estimation), and WinBUGs (Buckley, 2002).

In the following sections, we set out the basic model and fundamental equations. Then we deal with inference, goodness of fit, model selection, and diagnostics. Model interpretation is elaborated in conjunction with examples. Finally, we offer some comparisons between beta regression and reasonable alternatives along with criteria for deciding among them, and we note possible extensions and generalizations of beta regression. Practical issues regarding maximum-likelihood estimation and implementations in various statistical packages are discussed in the supplementary material referred to in Footnote 2.

Beta Regression Model and Maximum-Likelihood Estimation

Let Y be the dependent variable of interest and $i = 1, \ldots, N$, index the cases. Assume Y_i has a beta distribution with parameters ω and τ, that is,

$$f(y;\,\omega,\,\tau) = \frac{\Gamma(\omega+\tau)}{\Gamma(\omega)\Gamma(\tau)}\, y^{\omega-1}(1-y)^{\tau-1}, \tag{1}$$

where $y \in (0, 1)$, $\omega,\tau > 0$, and $\Gamma(\cdot)$ denotes the gamma function. Both ω and τ are shape parameters, with ω pulling density toward 0 and τ pulling density toward 1. The beta family includes a broad variety of distribution shapes. Two noteworthy cases are Beta(1, 1), which is equivalent to the uniform distribution over the unit interval, and the limit as $\omega, \tau \rightarrow 0$, which is a Bernoulli($p$), with p determined by the relative rate of convergence $\omega,\ \tau$ to 0. Figure 2 displays several examples; we encourage readers to plot others. The beta distribution reflects around .5 in the sense that $f(y;\,\omega,\tau) = f(1-y;\,\tau,\omega)$, so that reverse scoring beta random variables does not change estimates in a meaningful way. It has a mode only if both ω and τ are greater than 1, located at $(\omega - 1)/(\tau + \omega - 2)$. Further details about the beta distribution may be found in Gupta and Nadarajah (2004) or Johnson, Kotz, and Balakrishnan (1995). Weisstein (2005) is a particularly convenient resource, and Wolfram Research (2005) has a number of interactive demos.

Provided one is willing to assume the interval level of measurement, it is possible to transform any bounded scale in $[a, b]$ with known bounds into the unit interval without loss of generality using the transformation $y' = (y - a)/(b - a)$. In practice, this transformation usually needs to be modified slightly to avoid zeros and ones in the data (this issue is discussed in the supplementary material referred to in Footnote 2). We stress here that the bounds a and b refer to known theoretical minimum and maximum values on a

[2] Code to replicate the examples and graphics in this article can be downloaded from http://www.anu.edu.au/psychology/people/smithson/details/Index.html. A Readme file discusses practical estimation issues and tips. These materials also are accessible at http://dx.doi.org/10.1037/1082-989X.11.1.54.supp.

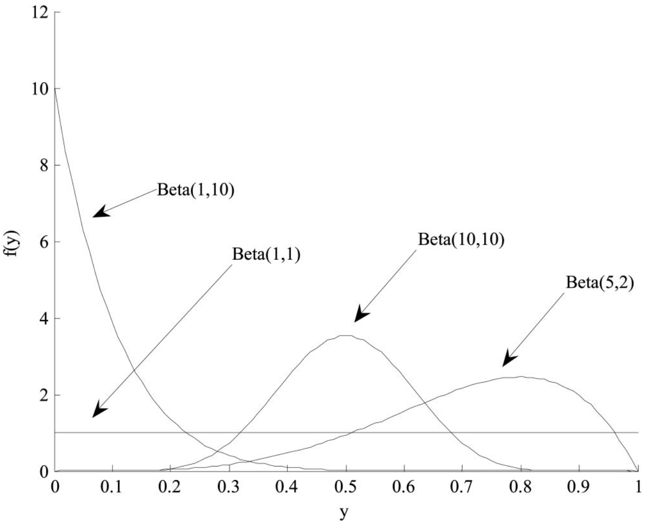


Figure 2. Beta densities.

scale, regardless of the smallest and largest values observed in data. The beta distribution defined over the unit interval sometimes is known as the *standard beta* to distinguish it from beta distributions defined over intervals with other bounds. We deal only with the standard beta in this article, and we also leave aside the problem of estimating scale endpoints (see Johnson et al., 1995, p. 221, for the standard approach to this problem in the univariate setting).

The usual (i.e., degrees of freedom) parameterization of the beta distribution is inconvenient for regression because ω and τ are both shape parameters, usually interpreted as prior sample sizes in Bayesian analysis of proportions (Pham-Gia, 1994). Unfortunately, shape parameters are difficult to interpret in terms of conditional expectations. The following reparameterization translates ω and τ into location and dispersion parameters. It is well-known that

$$E(Y) = \frac{\omega}{\omega + \tau}$$

and

$$Var(Y) = \frac{\omega\tau}{(\omega + \tau)^2(\omega + \tau + 1)} .$$

Let $\mu = \mathrm{E}(Y)$ and $\sigma^2 = \mathrm{Var}(Y)$. Simple algebra yields

$$\sigma^2 = \frac{\mu(1 - \mu)}{(\omega + \tau + 1)} .$$

Now, let $\phi = \omega + \tau$. Then we have

$$\omega = \mu\phi \tag{2}$$

and

$$\tau = \phi - \mu\phi. \tag{3}$$

The reparameterization results in a location parameter μ and a "precision" parameter ϕ—because Equations 2 and 3 imply $\sigma^2 = \mu(1 - \mu)/(\phi + 1)$, so the variance increases as ϕ decreases.

The variance is therefore a function of both the mean and the parameter ϕ. Because we propose modeling μ and ϕ separately, it is essential to bear in mind that although ϕ influences precision (and therefore dispersion), it is not the sole determinant of dispersion. A natural feature of a bounded random variable is that dispersion depends partly on location. However, note that μ and ϕ place no restrictions on each other, so they can be modeled separately.

The variance also must have an upper bound. As is readily seen from the formula

$$\sigma^2 = \frac{\mu(1 - \mu)}{(1 + \phi)} = \frac{\omega\tau}{(\omega + \tau)^2(\omega + \tau + 1)} ,$$

$\sigma^2 < 1/4$. The limit is approached when $\mu = 1/2$ as ϕ approaches 0. It is also the case that $\sigma^2 > 1/12$ when τ and ω both are < 1, and $\sigma^2 < 1/12$ when τ and ω both are > 1.

We use maximum-likelihood to estimate ω and τ, or equivalently μ and ϕ. Because the beta forms an exponential family, it also satisfies certain useful regularity conditions that help ensure maximum-likelihood estimates exist and are well-defined. For a technical discussion of the particular conditions required to ensure asymptotic properties of the maximum-likelihood estimators in generalized linear models, readers may consult Fahrmeier and Tutz (2001) and references therein, but it seems that beta regression generally satisfies these conditions. The log-likelihood for the ith observation y_i is

$$\ln L(\omega,\tau; y_i) = \ln \Gamma(\omega + \tau) - \ln \Gamma(\omega) - \ln \Gamma(\tau) + (\omega - 1)\ln(y_i) + (\tau - 1)\ln(1 - y_i). \tag{4}$$

The maximum-likelihood estimators are found in the usual way by maximizing the sum of the log-likelihoods over the observations. In general, this must be done numerically. It should be borne in mind that maximum-likelihood estimates are not necessarily unbiased, and the degree of bias will be greater for small samples (see Cribari-Neto & Vasconcellos, 2002, for an exploration of bias in such models).

To form a regression model in the extended generalized linear model (GLM) approach, we use two *link functions*, one for the location parameter μ and the other for the precision parameter ϕ. The link function is a nonlinear, smooth, monotonic function that maps the unbounded space of the linear predictor into the proper sample space of the observations, thereby "linking" the linear predictor with the observations. Let $\mathbf{X}$ and $\mathbf{W}$ be matrices of covariates (possibly overlapping or even identical), with $\mathbf{x}_i$ or $\mathbf{w}_i$ being the ith row vector from these matrices, and let β and δ be column vectors of coefficients. The GLM for the location parameter is the usual

$$f(\mu_i) = \mathbf{x}_i\beta,$$

where f is a monotonic, differentiable link function. The general form of the relationship between the means and variances is

$$\sigma_i^2 = v(\mu_i)u(\phi_i),$$

where v and u are nonnegative functions. The precision parameter ϕ_i is assumed to be modeled by

$$h(\phi_i) = \mathbf{w}_i\delta,$$

where h is another link function. Following Smyth's (1989) terminology, we refer to the likelihood function of $\boldsymbol{\beta}$ when δ is held constant as defining the *location submodel*. Likewise, the likelihood function of $\boldsymbol{\delta}$ when β is held constant defines the *dispersion submodel*.

For a beta-distributed dependent variable, the mean must lie in the open unit interval, so a link function that "squeezes" the real line into the unit interval is necessary. The logit link does this:

$$\ln\,[\mu_i/(1 - \mu_i)] = \mathbf{x}_i\beta,$$

because the logit, or log-odds, transformation $\ln\,[\mu/(1 - \mu)]$ maps a number $\mu \in (0, 1)$ onto the real line. The logit is the quantile function of the standard logistic distribution, which is why logistic regression—which also uses the logit link—has its name. Beta regression is, in a sense, a generalization of logistic regression when the dependent variable is a proportion. The logit link is desirable from an interpretation standpoint because the resulting regression coefficients can be interpreted as log-odds. Cox (1996) considered various link functions for regression models of continuous proportions in a quasi-likelihood framework and found that the logit performed well.

In this article we restrict attention to the logit link, but other link functions can be used. Any link function whose domain is the unit interval (e.g., an inverse cumulative distribution function) is possible. The probit, cauchit, and complementary log-log links are common examples. The probit link, $\Phi^{-1}(\mu)$, rescales the coefficients in terms of the standard normal distribution, which may be preferred to log-odds by users more familiar with effect size measures such as Cohen's *d*. Up to rescaling, estimates will be nearly identical to logit, but the probit link is more susceptible to changes in the tails. The cauchit link, $\tan[\pi(\mu - .5)]$, the inverse cumulative distribution function (CDF) of the standard Cauchy distribution, is very useful if outliers in the space of the linear predictor, that is, high leverage points, are suspected. The Cauchy distribution is very heavy tailed and makes less extreme predictions for the expected value of the dependent variable than the probit or logit for large values of the linear predictor. The complementary log-log link, $\ln\,[-\ln\,(1 - \mu)]$ is asymmetric and is useful in certain applications; it scales the coefficients in terms of the hazard function rather than the log-odds. In the absence of theoretical considerations, a link function is chosen to provide a simple relationship between the covariates and the dependent variable. Collett (2003) provided extensive discussion of link functions for binary data, most of which extends naturally to beta regression.[3] The identity link, as used by Shazam 7, is generally not preferred for bounded variables because it makes out-of-bounds predictions and does not incorporate the general notion of diminishing returns for observations near the bounds.

The precision parameter ϕ must be strictly positive because a variance cannot be negative. The log link has this property:

$$\ln(\phi_i) = -\mathbf{w}_i\delta.$$

We use the negative sign to make the interpretation of the coefficients δ easier. Because ϕ is a precision parameter, a positive-signed δ_j indicates smaller variance, which is potentially confusing. It seems more natural to model dispersion rather than precision, and the negative sign enables us to do so. This is the difference between our parameterization and those in the literature on beta regression cited so far. Unlike the location submodel, there does not seem to be any other obviously useful link for the dispersion.

Inverting the link functions to give predicted values, the location submodel may be written as

$$\mu_i = \frac{\exp(\mathbf{x}_i\beta)}{1 + \exp(\mathbf{x}_i\beta)}, \tag{5}$$

and the dispersion submodel is

$$\phi_i = \exp(-\mathbf{w}_i\delta). \tag{6}$$

Note that we have incorporated the intercepts for both models into the coefficient vectors. We shall adopt the convention here that the first elements in $\mathbf{X}$ and $\mathbf{W}$ are $x_{0i} = w_{0i} = 1$, and the first elements in β and δ are β_0 and δ_0, the submodel intercepts. The reparameterized log-likelihood kernel for the regressors is

[3] One further possibility that we mention is to use a generalized link function. The idea is to treat the link function as a nuisance parameter and then estimate it rather than fixing it a priori. For instance, the Aranda–Ordaz class embeds the complementary log-log and logit links in a one-parameter family. It is then possible to estimate this parameter along with the rest of the model. The analogy to Box–Cox transformations is obvious, and this strategy shares many of the flaws, in particular the potential for capitalization on chance and a drastic loss of efficiency. See Collett (2003) for a discussion and citations to the original literature.

$$\begin{aligned}\ln L(\beta,\delta;\ \mathbf{y}_i,\mathbf{X},\mathbf{W}) &= \ln\Gamma[(e^{\mathbf{X}\beta-\mathbf{W}\delta} + e^{-\mathbf{W}\delta})/(1 + e^{\mathbf{X}\beta})] \\ &\quad -\ln\Gamma[e^{\mathbf{X}\beta-\mathbf{W}\delta}/(1 + e^{\mathbf{X}\beta})] - \ln\Gamma[e^{-\mathbf{W}\delta}/(1 + e^{\mathbf{X}\beta})] \\ &\quad + [e^{\mathbf{X}\beta-\mathbf{W}\delta}/(1 + e^{\mathbf{X}\beta}) - 1]\ln(y) \\ &\quad + [e^{-\mathbf{W}\delta}/(1 + e^{\mathbf{X}\beta}) - 1]\ln(1 - y).\end{aligned}$$

The score function (gradient of the likelihood) and the Hessian (matrix of second partials) can be obtained explicitly in terms of the polygamma function, so asymptotic standard-error estimates can also be computed numerically.

Inference, Goodness of Fit, and Diagnostics

As mentioned earlier, because the estimation is by maximum-likelihood, the usual inferential machinery of Wald statistics, likelihood ratio tests, and Lagrange multiplier (score) tests is available. We also recommend resampling or permutation as methods that do not make assumptions about asymptotic properties of the model. At present the small sample properties of beta regression are generally unknown. Parameter estimates are consistent but not unbiased, and in small samples the influence of bias is strongest (of course). A few hundred observations seem to be sufficient for practical infinity, and we have fit reasonable examples with sample sizes as small as 10, but further studies are obviously necessary to quantify the amount of bias. Model comparison can be done via the likelihood ratio test, using twice the difference between the log-likelihoods of a full model and a restricted model whose covariates are a subset of the full one. Information criteria such as Akaike's information criterion (AIC) or the Bayesian information criterion (BIC) also can be used for this purpose. Both quantities are penalized chi-square values, with the penalties defined by the number of model parameters, *k*, and, in the case of the BIC, the number of observations, *N*. The AIC is commonly defined as

$$AIC = -2\ln L_{\text{fit}} + 2k.$$

AIC has the well-known flaw of being dependent on sample size; it tends to capitalize on chance and thus favor more complex models when the sample size is large. For this reason, many prefer the BIC. There are a number of ways to define this statistic, but we use

$$BIC = -2\ln L_{\text{fit}} + 2k\ln N$$

in the examples below. In general, the BIC tends to favor simpler models than AIC or the likelihood ratio test when *N* is large because the change in likelihood for the addition of an additional parameter needs to be large enough to overcome the penalty of ln *N*, which grows as the sample size does. (One should not rely overly much on any of these criteria, however.)

To assess the global goodness of fit of the model, it would be useful to have an analog to multiple-R^2 in normal-theory ordinary least squares (OLS) regression. Ferrari and Cribari-Neto (2004) considered the correlation between observed and predicted values as a basis for a measure of goodness of fit. However, this statistic does not take into account the effect of dispersion covariates, and thus its utility is limited. More broadly, generalizing the R^2 to contexts beyond the linear model has not proven to be an easy task. Conceptually different approaches that lead to the same thing in the normal theory do not lead to the same thing outside it. This issue is reviewed thoroughly in Long (1997) in the context of the generalized linear model, and we recommend the discussion therein.

A simple candidate is a proportional reduction of error (PRE) statistic based on log-likelihoods, which is McFadden's (1974) pseudo-R^2:

$$\text{PRE} = 1 - \ln L_{\text{null}}/\ln L_{\text{fit}},$$

where ln L_{null} is the log-likelihood of the null model as defined earlier, and ln L_{fit} is the log-likelihood of whatever model has been fitted. If it is desirable to consider other reference models, the PRE approach can easily be adapted by substituting a different model for the null model. The PRE statistic is often disappointingly small and may not attain the theoretical maximum value of 1; thus, many different adjustments to PRE have been proposed. For instance, Nagelkerke's (1991) statistic is adjusted by the maximum possible PRE in the sample. Its formula is

$$\text{PRE}_{\text{Nagelkerke}} = [1 - (L_{\text{null}}/L_{\text{fit}})^{2n}]/[1 - (L_{\text{null}})^{2n}],$$

where L_{null} and L_{fit} are the likelihoods (*not* log-likelihoods) of the null and fitted models, respectively.

Diagnostics require suitable influence measures and assessment of the residuals. Ferrari and Cribari-Neto (2004) provided a derivation of deviance residuals (i.e., based on log-likelihood contributions), but their model does not include dispersion covariates and it is unclear what their expression for deviance residuals means when dispersion covariates are present. To assess influence of a case on the estimated coefficients, we advocate leave-one-out jackknifing after plotting to see if particular cases seem to have unusually large residuals relative to the rest. Simply plotting the predicted values versus the raw residuals or plotting the sorted predicted values versus the corresponding observed values is highly informative as ill-fitting cases usually jump out visually. Another effective way to screen for high leverage points is to estimate the model using the probit, logit, and cauchit links and see whether the estimated coefficients—after rescaling them to lie on a common range so the coefficients are comparable—change substantially for different links. The probit is most sensitive to change in the linear predictor, and the cauchit is least sensitive. The logit falls in between, which makes it a good general-purpose compromise. As with ordinary GLMs, reestimation after deleting a suspect case gives a useful sense of its influence on the coefficients in both submodels. Beta models fit quite rapidly on modern computers so multiple model refits are feasible. Collett (2003) has a thorough discussion of many aspects of model fit.

One commonly used diagnostic in ordinary regression is *not* advisable here, namely, screening the dependent variable to see if its marginal distribution is beta. The beta regression model posits a *conditional* beta distribution. That is, it claims $Y_i|(\mathbf{X}_i,\mathbf{W}_i) \sim \text{Beta}[\omega(\mathbf{X}_i,\mathbf{W}_i),\tau(\mathbf{X}_i,\mathbf{W}_i)]$, which does not imply that $Y_i \sim \text{Beta}(\omega,\tau)$ *unconditionally*. In fact, beta regression can be applied fruitfully to dependent variables whose marginal distributions are quite far from beta. Although we note that plotting the dependent variable marginally is not helpful, it is useful to plot conditional densities. If they do not fit a beta distribution well, then that is a sign that the model may not be appropriate. We illustrate this point in the following examples.

Interpreting and Evaluating the Location and Dispersion Submodels: Examples

This section deals with the evaluation and interpretation of predictor effects in the location and dispersion submodels. Many readers will find links between this material and the GLM, especially logistic regression and log-linear models, because the location submodel is linear in the logit scale and the dispersion submodel is linear in the log scale. We present three illustrative examples. The first is from a simple experiment with a 2 × 2 factorial design, and the other two are based on observational studies.

Example 1: 2 × 2 Factorial Design

Deady (2004) studied naïve mock-jurors' responses to the conventional two-option verdict (guilt vs. acquittal) versus a three-option verdict setup (the third option was the Scottish "not proven" alternative). She crossed this between-subjects factor with another between-subjects two-level factor in which there was conflicting testimonial evidence (conflict condition) versus one in which there was no conflicting evidence (no-conflict condition). Participants were 104 first-year psychology students at The Australian National University. One of the dependent variables was the juror's degree of confidence in her or his verdict, expressed as a percentage rating (0–100). Deady hypothesized that the conflicting evidence would lower juror confidence and that the three-option condition would increase it.

Figure 3 displays histograms of the confidence ratings for each of the four experimental conditions. At least one of the graphs shows strong skew, and there appears to be heteroscedasticity as well. One notable feature is a response bias for .5, particularly in the two-option cases, and for .75 in all conditions. We have also superimposed the predicted distributions from the final beta regression model on these pictures, which is discussed more fully below. For illustration, we use the previously mentioned rescaling of the dependent variable to lie in the (0, 1) interval (effectively shrinking the interval to [.005, .995] to avoid zeros and ones):

$$y'' = [y'(N-1) + 1/2]/N = [(y/100)(103) + 1/2]/104.$$

The experimental condition design matrix uses effects coding so that no conflict = −1 and conflict = 1, whereas the two-option verdict = −1 and the three-option verdict = 1. A 2 × 2 ANOVA with Y as the dependent variable yields a nonsignificant model overall, $F(3, 100) = 1.797$, $p = .153$, and a significant result for Levene's homogeneity test, $F(3, 100) = 3.394$, $p = .021$. A 2 × 2 ANOVA using the logit transform of Y also yields a nonsignificant model, $F(3, 100) = 1.540$, $p = .209$. None of the individual coefficients reaches significance in either model.

Turning now to a beta regression approach, for the null model $\ln L_{\text{null}} = 28.200$. In a result reminiscent of the two OLS models mentioned above, a beta regression model with verdict, conflict, and their interaction term entered into the location submodel only yields $\ln L_{\text{fit}} = 30.558$, so the chi-square change is 4.716, not significant with 3 *df*. However, including verdict, conflict, and their interaction term in the dispersion submodel as well produces $\ln L_{\text{fit}} = 40.095$, so the chi-square change is 23.790 with 6 *df* ($p < .001$). The overall model effect size is $PRE = 1 - (\ln L_{\text{null}}/\ln L_{\text{fit}}) = 1 - (28.2/40.095) = .297$.

All of the terms in the dispersion submodel are significant. Moreover, the resultant model reveals a significant interaction effect in the location submodel (and a nearly significant main effect for conflict). This example illustrates how new terms in one submodel can clarify the effects in the other. Table 1 displays the model coefficients, asymptotic standard error estimates (using the Hessian, also known as the "information matrix"; see any text on maximum-likelihood estimation, e.g., Rose & Smith, 2002, chapter 12), and significance tests.

In the location submodel, predicted means m_{ij} are computed in the usual fashion with the inverse-link function after the coefficients have been substituted into the model.[4] For instance, in the no-conflict/two-option cell we have

$$\ln[m_{11}/(1-m_{11})] = b_0 - b_1 + b_2 - b_3 = 0.9120 - 0.0050 + 0.1685 - 0.2800 = 0.7955,$$

so the predicted cell mean is

$$m_{11} = \exp(0.7955)/[1 + \exp(0.7955)] = .6890.$$

[4] Many software packages allow computation of linear and nonlinear functions of estimated parameters. We recommend using these commands when they are available because they usually compute proper standard errors along with the desired estimate.

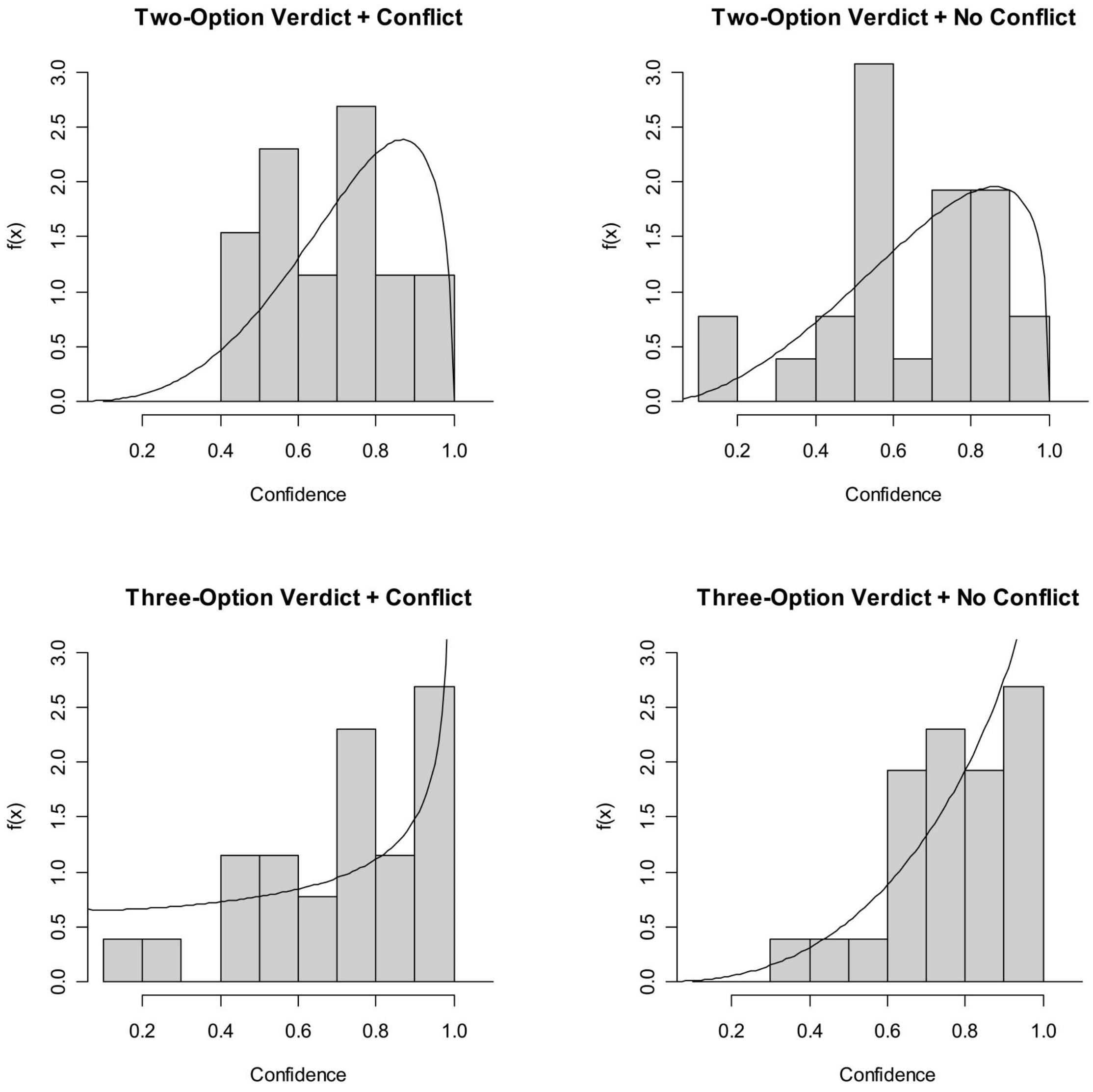


Figure 3. Confidence ratings for four experimental conditions.

The cell coefficients, predicted logits, predicted means, and observed means are shown in Table 2. The cell coefficients are the estimated interaction terms, with the main effects terms as row and column entries. The coefficients in the location submodel may be interpreted in terms similar to odds ratios. For instance, mean confidence is higher in the no-conflict than the conflict condition. The relevant coefficient, $b_2 = 0.1685$, is the difference between the logit of $m_{1.}$ and the logit of the grand mean $m_{..}$: $1.0805 - 0.9120 = 0.1685$. Therefore, $\exp(b_2)$ is the ratio $[m_{1.}/(1 - m_{1.})]/[m_{..}/(1 - m_{..})] = [.7466(1 - .7466)]/[.7134(1 - .7134)] = 1.1836 = \exp(0.1685)$.

In the dispersion submodel, predicted variances use both predicted means and predicted dispersion parameters. Table 3 shows the dispersion submodel coefficients, dispersion parameter estimates, and the predicted and observed variances. Denoting the sample estimate of ϕ_{ij} by f_{ij}, recall that the *negative* coefficients in a cell, its row, and column all sum to $\ln(f_{ij})$. For instance, for the two-option/no-conflict cell we have

$$\ln(f_{21}) = -d_0 + d_1 - d_2 + d_3 = 1.1737 + 0.3300 + 0.2195 - 0.3162 = 1.4071.$$

Substituting the dispersion and mean estimates into Equation 2 for the variance yields

$$s_{21}^2 = [.6890(1 - .6890)]/[1 + \exp(1.4071)] = 0.0421.$$

A natural way to evaluate deviations of the cell dispersion parameters away from the grand mean dispersion parameter is by taking ratios. For instance, f_{ij} is higher in the conflict condition than in the no-conflict condition. The relevant coefficient, $d_2 = 0.2195$, is the difference between $\ln(f_{2.})$

Table 1
Full Model Coefficients, Standard Errors, and Significance Tests For Example 1

Parameter	Coefficient	*SE*	*p*
Location submodel			
b_0	0.9120	0.1040	.0000
b_1 (verdict)	0.0050	0.1040	.4808
b_2 (conflict)	0.1685	0.1040	.0525
b_3 (Verdict × Conflict)	0.2800	0.1040	.0035
Dispersion submodel			
d_0	−1.1737	0.1278	.0000
d_1 (verdict)	0.3300	0.1278	.0049
d_2 (conflict)	−0.2195	0.1278	.0429
d_3 (Verdict × Conflict)	−0.3162	0.1278	.0067

and ln $(f_{..})$: 1.1737 − 0.9542 = 0.2195. Likewise, cell variances may be compared with the grand mean variance via ratios.

We are now in a position to interpret the entire model. From Tables 2 and 3 we can see that mean confidence levels are highest in the no-conflict/three-option and the conflict/two-option conditions, but the conflict/three-option condition not only has a lower mean but also has a substantially larger variance than the other three conditions. There are main effects on variance for both conflict and verdict options, with the conflict and three-option conditions having the greater variance. Inspection of Table 3 suggests that these effects are again driven by the larger variance in the conflict/three-option condition. Looking at the predicted distributions shown in Figure 3, we can see that much of the misfit in the model is due to the .5/.75 response bias mentioned previously—otherwise the predicted densities track the histograms well. The effect of the bias is to create a spike at these points. This sort of pattern is not uncommon in a scale of the type used here to assess confidence and is often a sign of a noninformative response from the participant. One way to handle this problem formally would be to use a finite mixture model, where the response scale is decomposed into two parts, a nonresponse and a genuine response. We are working on procedures for estimating such models in the beta regression context, though in a relatively small data set it is unlikely to be possible to estimate a finite mixture model.

Table 2
Location Submodel Coefficients, Logits, and Predicted Means For Example 1

Statistic	2-option verdict	3-option verdict	Total
Coefficients[a]			
No conflict	−0.2800	0.2800	0.1685
Conflict	0.2800	−0.2800	−0.1685
Total	−0.0050	0.0050	0.0000
Logit(m_{ij})			
No conflict	0.7955	1.3656	1.0805
Conflict	1.0185	0.4685	0.7435
Total	0.9070	0.9170	0.9120
Predicted means m_{ij}			
No conflict	.6890	.7967	.7466
Conflict	.7347	.6150	.6778
Total	.7124	.7144	.7134
Observed means			
No conflict	.6885	.7908	.7406
Conflict	.7228	.6643	.6941
Total	.7056	.7300	.7178

[a] Cell coefficients correspond to the interaction effect parameter b_3, column coefficients to the main effect parameter b_1, and row coefficients to the main effect parameter b_2 in Table 1.

Table 3
Dispersion Submodel Coefficients, Log-F Estimates, and Predicted Variances for Example 1

Statistic	2-option verdict	3-option verdict	Total
Coefficients[a]			
No conflict	0.3162	−0.3162	−0.2195
Conflict	−0.3162	0.3162	0.2195
Total	−0.3300	0.3300	0.0000
ln(f_{ij})			
No conflict	1.4071	1.3794	1.3932
Conflict	1.6004	0.3079	0.9542
Total	1.5037	0.8437	1.1737
Predicted $s_{ij}{}^2$			
No conflict	.0421	.0326	.0376
Conflict	.0327	.1003	.0607
Total	.0373	.0614	.0483
Observed $s_{ij}{}^2$			
No conflict	.0408	.0295	.0370
Conflict	.0244	.0843	.0535
Total	.0322	.0588	.0452

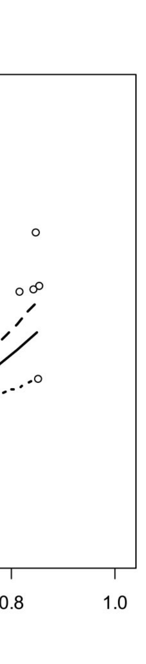
[a] Cell coefficients correspond to the interaction effect parameter d_3, column coefficients to the main effect parameter d_1, and row coefficients to the main effect parameter d_2 in Table 1.

More generally, Long (1997, pp. 61–82) provided a detailed and highly readable examination of methods useful in logistic regression and other generalized linear models. These extend naturally to the beta regression case and should be given serious consideration. As he noted regarding binary response models (BRMs, e.g., logistic or probit regression), "Since the BRM is nonlinear, no single approach to interpretation can fully describe the relationship between a variable and the outcome probability. You should search for an elegant and concise way to summarize the results that does justice to the complexities of the nonlinear model" (Long, 1997, p. 61). One of the most useful ways to understand how the dependent variable changes with a covariate is to plot the predicted values over the range of the selected covariate, holding all other covariates at constant values, for example, their means or some other relevant baseline. In an experimental design such as the one discussed above, it is straightforward to plot the conditional densities predicted by the model, as shown in Figure 3. R code to generate the plots is available on the Web pages referred to in Footnote 2.

Example 2: Stress, Depression, and Anxiety

This example illustrates the use of a continuous predictor. The unpublished data come from a sample of 166 nonclinical women in Townsville, Queensland, Australia. The variables are linearly transformed scales from the Depression Anxiety Stress Scales (Lovibond & Lovibond, 1995), which normally range from 0 to 42. Figure 4 shows kernel density estimates for the two variables.

As one would expect in a nonclinical population, there is an active floor for each variable, with this being most pronounced for anxiety. It should be clear that anxiety is strongly skewed. In Figure 5, a local linear regression (lowess) curve is plotted for the prediction of Anxiety × Stress (Cleveland, 1993, Chap. 3). Lowess uses a nonparametric smoothing approach to find a good-fitting curve that tracks the data well. It thus provides a relatively unbiased view of the conditional location relationship in the data without the restrictions of a globally fit parametric model. The lowess curve clearly suggests a nonlinear relationship between anxiety and stress given its hockey-stick shape. Likewise, heteroscedasticity is to be expected here on theoretical as well as empirical grounds: It is likely that someone experiencing anxiety will also experience a great deal of stress, but the converse is not true, because many people might experience stress without being anxious. That is, stress could be thought of as a necessary but not sufficient condition for anxiety. If so, the relationship between the variables is not one-to-one plus noise, as predicted by ordinary homoscedastic regression, but instead is many-to-one plus noise, which predicts heteroscedasticity (more detailed discussions of tests of necessity vs. sufficiency may be found in Smithson, 2005, and Smithson and Verkuilen, 2006).



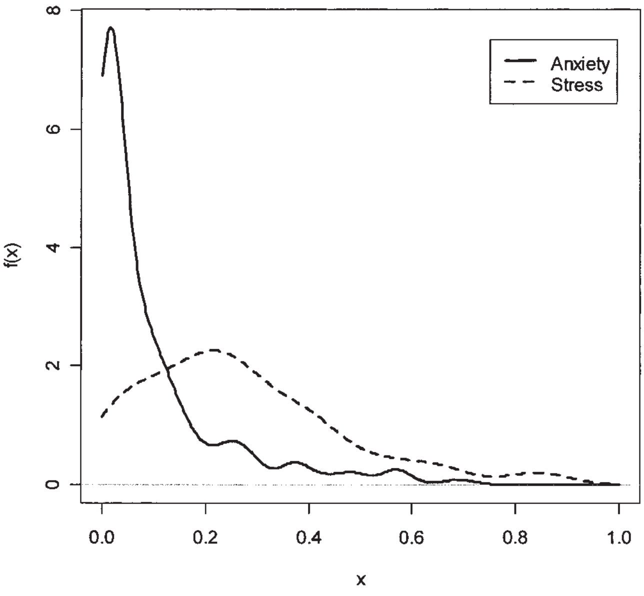


Figure 4. Kernel density plots of anxiety and stress.

We fit a beta regression with stress as a predictor in both the location and dispersion models:

$$\ln[\mu/(1-\mu)] = \beta_0 + \beta_1\text{Stress}$$

$$\ln\phi = -\delta_0 - \delta_1\text{Stress}$$

Results from these regressions are included in Table 4. All coefficients in both the location and dispersion submodels are significant, indicating that stress predicts both the location and heteroscedasticity, with higher values of stress associated with increased variability in anxiety.

How well does the beta regression track the data? Figure 5 shows the scatterplot, the lowess curve, and two predicted-value curves, one from the beta regression model and the other from an OLS model. If the two models are not sufficiently different and there is no other basis for choice, the simpler linear one is obviously preferable. However, the beta predicted values do track the point cloud better than the linear regression. The OLS model BIC = −2*157.887 + 3* ln (166) = −300.438, whereas the beta regression BIC = −2*301.960 + 4* ln (166) = −583.472. For anxiety, the beta predicted curve strongly resembles the lowess curve in Figure 5, but unlike the lowess curve, the beta model has only a few parameters. If one wants to make comparisons with other data, this is helpful.

Table 4
Regression Coefficients and Summary Statistics for Example 2

	Null model		Stress model	
Parameter or statistic	Coefficient	*SE*	Coefficient	*SE*
Location submodel				
b_0 (anxiety)	−2.2440	0.0988	−4.0237	0.1442
b_1			4.9414	0.4409
Dispersion submodel				
d_0 (anxiety)	−1.7956	0.1230	−3.9608	0.2511
d_1			4.2733	0.7532
Summary statistics for anxiety				
−2 ln L	−478.9		−603.9	
PRE			0.207	
AIC	−474.9		−595.9	
Pseudo-R^2			0.569	

Note. PRE = proportional reduction of error; AIC = Akaike's information criterion.

Another possible OLS model is to add a quadratic term. Doing so yields a model with BIC = −2*170.240 + 4* ln (166) = −320.032, which still is not nearly as good as the beta regression model. Note that fit statistics alone do not provide an adequate assessment of a model's performance. The quadratic model still does not handle heteroscedasticity. In addition, at the floor of stress (which in this sample comprises about 10% of the data), both the linear and quadratic OLS models make out-of-sample predictions for anxiety, whereas beta regression never makes out-of-sample predictions (by construction). Finally, there is little rationale for adding the quadratic term, whereas we might expect on theoretical grounds that there would be heteroscedasticity.

Instead of adding more terms to the OLS model, we could attempt to stabilize the variance and linearize the relationship by transforming the original Anxiety scale. A popular approach is the "ladder of powers" (Tukey, 1977), a family of transformations of the form $y^* = y^{(1 - b)}$, where b takes values from {-2, −1, 0, 0.5, 1, 1.5, 2} with the case $b = 1$ denoting the natural log transformation. Typically, b is estimated via a linear regression of the log of the absolute value of the OLS residuals as predicted by the log of the predictor. In this example, the estimate of b is 0.109 with a standard error of 0.042, suggesting that $b = 0$ or perhaps $b = 0.5$ would be the most reasonable candidates. However, no value of b from 0 to 0.5 succeeds in stabilizing variance (see also the automated transformation procedure applied to this example later in this article), nor do transformations corresponding to other nearby values of b, such as the log. The main reason for this is that the floor of anxiety is also its mode, and the cases at the floor cover nearly the same range of values on stress as the rest of the data.

Figure 6 shows the predicted standard error of estimate for the beta regression model and the untransformed OLS model. The general "football" shaped standard error function for the beta model is plotted along with the OLS standard error of estimate. The football is shifted toward the top end of the Stress scale by the positive dispersion coefficient. Some misfit in the model is evident from the relatively large residuals near the bottom of the Stress scale.

Example 3: Sequential Regression With Dyslexia and IQ Predicting Reading Accuracy

This example presents a simple sequential beta regression using Pammer and Kevan's (2004) previously mentioned comparison between children with and without dyslexia. They began by citing the suggestion that the reading performance of dyslexic readers compared with normal readers may simply be the result of differences in general cognitive ability. A test of this suggestion is to take normal and

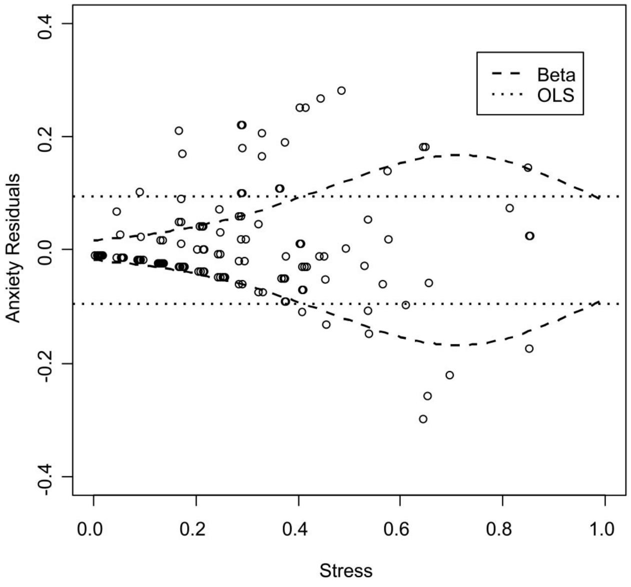


Figure 6. Standard error plots for linear regression versus beta regression. OLS = ordinary least squares.

dyslexic readers who differ on IQ, where IQ would therefore be expected to explain reading ability, and ascertain whether there still is anything about dyslexia that predicts reading skill, even when IQ is taken into account. As mentioned at the beginning of this article, such a test is complicated by the heavily skewed reading-accuracy scores among normal readers—with many participants achieving the maximum possible score—and heterogeneity of variance between the two groups.

Participants were recruited from primary schools in the Australian Capital Territory, resulting in a sample of 44 children (19 children with dyslexia and 25 controls). They ranged in age from 8 years 5 months to 12 years 3 months, with an average age of 10 years 6 months. Consultation with specialist teachers indicated that none of the children had any history of neurological illness or emotional problems and that all had experienced normal educational opportunities. All children had normal to corrected-to-normal visual acuity and normal hearing.

The following analyses use nonverbal IQ scores converted to z scores and dyslexia status coded as -1 = nondyslexic and 1 = dyslexic. Now we use a beta regression approach. Entering IQ first in both the location and dispersion submodels yields a chi-square change from the null model of $\chi^2(2) = 2*(34.9238 - 26.4206) = 17.0064$, $p = .0002$. The location submodel coefficient for IQ is positive as expected and significant, whereas the dispersion submodel coefficient is not significant.

Entering the main effect term for dyslexia status results in a large chi-square increase over the IQ-only model, $\chi^2(2) = 2*(61.2569 - 34.9238) = 52.6662$, $p < .0001$. Moreover, this model is significantly improved by adding the interaction term to the location submodel—the chi-square change is $\chi^2(1) = 2*(65.9019 - 61.2569) = 9.4620$, $p = .0237$—whereas adding the interaction term to the dispersion submodel yields only negligible change in model fit.

The coefficients and significance tests for our final model are shown in Table 5 and compared with the results from an OLS regression model using the logit transform of accuracy scores. Unlike the OLS model, in the beta regression model it is clear that dyslexia status accounts for both location and dispersion when the effects of IQ have already been taken into account, and IQ still makes an independent contribution. Although the OLS model's coefficients share the same signs with the beta regression location submodel coefficients, their magnitudes differ substantially. Moreover, the standard errors in the OLS model are inflated because of its inability to separately model heteroscedasticity. The OLS model can say nothing about the effects of IQ or dyslexia status on the variability in accuracy scores.

Both submodels are readily interpretable in ways that make sense out of the data. In the location submodel, the control group logit is $\exp(b_1) = 2.0995$ times greater than the grand-mean logit, or $\exp(2b_1) = 4.4079$ times greater than the dyslexic group logit. An IQ score one standard deviation above the mean predicts a logit that is $\exp(b_2) = 1.6263$ times higher than the grand-mean logit. The interpretation of the main effects is that accuracy declines for the dyslexic group and increases with IQ. The interaction effect indicates that the positive relationship between IQ and accuracy holds for the nondyslexic group ($b_2 - b_3 = 1.0676$) but not for the dyslexic group ($b_2 + b_3 = -0.0949$), a finding that makes clinical sense. Dyslexic readers have difficulty reading regardless of their general cognitive ability, whereas cognitive ability predicts reading accuracy for nondyslexics. The predicted means are displayed in the upper half of Table 6.

Table 5
Model Coefficients, Standard Errors, and Significance Tests for Example 3

Parameter	Coefficient	*SE*	*p*
OLS regression (logit transform)			
b_0	1.6011	0.2259	.0000
b_1 (dyslexia)	−1.2056	0.2259	.0000
b_2 (IQ)	0.3954	0.2255	.1187
b_3 (Dyslexia × IQ)	−0.4229	0.2255	.0681
Beta regression			
Location submodel			
b_0	1.1232	0.1509	.0000
b_1 (dyslexia)	−0.7417	0.1516	.0000
b_2 (IQ)	0.4863	0.1671	.0018
b_3 (Dyslexia × IQ)	−0.5812	0.1726	.0003
Dispersion submodel			
d_0	−3.3044	0.2265	.0000
d_1 (dyslexia)	−1.7465	0.2940	.0000
d_2 (IQ)	−1.2290	0.4596	.0037

Both coefficients in the dispersion model are negative, indicating underdispersion for the dyslexic group and higher IQ children. Again, it is worth bearing in mind that the decrease in variance for higher IQ children is greater than the decrease that would be expected solely from the effect of higher accuracy scores. Likewise, a possible explanation for the lower variance in the dyslexic group stems from the location model interaction effect described above; dyslexia thwarts the effect of IQ on reading accuracy.

Both the effect of group and the effect of IQ seem large relative to the achievable range of the variance. The predicted variances displayed in Table 6 bear out this impression, showing values ranging from .0005 to .0896, covering a large part of the 0–.25 range for variances of beta-distributed variables. Figure 1 displays the predicted distributions generated by this model for the controls versus dyslexics superimposed on the histogram, illustrating how effective the beta regression model is in modeling skew as well as location and dispersion.

Table 6
Predicted Means and Variances for Dyslexic–IQ Model (Example 3)

IQ level	Control	Dyslexic	Total
Predicted means(m_{ij})			
IQ = 1 *SD* above mean	.9494	.5712	.8334
IQ = 1 *SD* below mean	.6894	.6169	.6540
Total	.8659	.5943	.7546
Predicted variance ($s_{ij}{}^2$)			
IQ = 1 *SD* above mean	.0028	.0005	.0015
IQ = 1 *SD* below mean	.0896	.0051	.0252
Total	.0202	.0015	.0066

Conclusion

Our title “A Better Lemon Squeezer?” refers to the fact that badly skewed heteroscedastic data, and even multimodal data, arise in practice frequently. These are the lemons of research in the social and behavioral sciences, and when you are handed lemons, it is best to learn how to make lemonade rather than wishing (or pretending) you had something else. Instead, far too often a suspect model is applied or some mostly ineffective remedial measures are taken in the hopes that the problems will go away. Even worse, data are simply discarded as being not analyzable. Finally, investigators frequently “deskew by design,” dropping potentially relevant and informative dependent variables from consideration simply because they do not look sufficiently normal.

There are several alternative approaches to dealing with the combination of skewness and heteroscedasticity in variables with scales bounded at both ends. We agree with Kieschnick and McCullogh’s (2003, pp. 196–197) assertion that the most common practices are seriously flawed. OLS linear regression assumes that the dependent variable covers the real line and is conditionally normally distributed, and the conditional expectation function is linear. Clearly none of these assumptions hold for bounded dependent variables. Moreover, boundedness implies that the conditional variance depends on the mean. However, several alternatives for dealing with skewness and heteroscedasticity deserve serious consideration along with beta regression. We consider the following potentially viable alternatives:

1. ordinal regression,
2. robust regression,
3. OLS regression using a transformation and normal error term,
4. Tobit regression with censoring,
5. alternative parametric models, and
6. a semi-parametric quasi-likelihood approach.

Although it is beyond the scope of this article to treat these in detail, we summarily compare them with beta regression and suggest criteria for choosing among them.

Ordinal regression models have been available for several decades now, and programs to estimate them are found in most mainstream statistics packages. They are straightforward extensions of the category thresholds approach to binary logistic or probit regression to discrete ordinal data. We refer readers to Long (1997) for a basic introduction to ordinal models. Although they are commonly used in other fields, they seem relatively unknown in psychology, except for confidence ratings used in signal detection experiments. In general, relatively coarse discrete data, such as four- or five-option Likert scales, can be fruitfully analyzed using this approach, but estimation breaks down when the number of response categories grows large unless the sample also is very large. If the response scale is coarse, ordinal regression should be considered over beta regression. The standard ordinal regression model does not handle heteroscedasticity, but some programs can estimate heteroscedastic models, for example, SPSS PLUM, SAS nlmixed, or Stata ologit, oprobit, and slogit. Beta regression and ordinal regression are similar in approach, and exploring their similarities is an active line of research for us.

Robust regression refers to techniques initially developed by Huber and Hampel and summarized in Huber (1981) and Hampel, Ronchetti, Rousseeuw, and Stahel (1986). A number of variants have been proposed more recently. Stautde and Sheather (1990) or Wilcox (2005) are more accessible references. These approaches originally were intended to handle problems with estimators such as the mean or standard deviation posed by outliers. Roughly speaking, the idea behind robust estimators is to estimate parameters using some kind of trimming or down-weighting of the most extreme observations. The model of the data behind robust regression differs from beta regression in two important respects. First, it usually starts with the premise that the dependent variable is unbounded and the underlying distribution is continuous, often Gaussian or at least symmetrical, but contaminated by outliers. (However, proponents such as Wilcox, 1998, claim that even when there are no outliers and distributions are skewed, robust regression offers a substantial advantage over standard techniques.) This is not the case with bounded scales, where skew and heteroscedasticity are directly attributable to the bounds and variance is always finite.

The second difference involves the relationship between linear predictor and dependent variable in a bounded space. There are diminishing returns in changes in scale near the boundaries of the space, which induces positive curvature (Schönemann, 1983). This aspect is not something built into the robust regression model, as far as we know. Although polynomial terms can be used to handle this curvature, they do not respect the bounds on the domain, only approximate the diminishing returns phenomenon, generally lack any theoretical motivation (the best justification for polynomial regression is usually from the perspective of a Taylor series approximation to a nonlinear function), and can generate absurd predictions. A related point is that a standard robust regression model does not address heteroscedasticity per se, whereas the beta regression model explicitly does so. There is definitely a place for robust methods, but in our experience it rarely makes much of a difference on bounded scale data.

A third popular practice is transforming the dependent and/or independent variable and then applying OLS regression to the transformed variable(s). We have already described the ladder-of-powers approach and applied it to our second example. Such transformations can produce linearity and stabilize variance. However, they may fail in the latter. Another crucial test of the adequacy of this approach is whether the residuals are normally distributed as assumed. The transformed OLS regression model fails on both counts in our examples and will generally do so in the presence of an active floor or ceiling. Transformations of this kind are most likely to be effective if there is no active floor or ceiling in the dependent variable and if the log of the absolute residuals is linearly related to the log of the predictor. One important advantage that beta regression shares with other GLMs over the ladder-of-powers transformations stems from the fact that the transformations transform raw data, whereas the link function in any GLM transforms expected values. The error distributions in the GLMs are more likely to be well-behaved as a consequence. However, researchers understandably may want to explore this transformation option before embarking on beta regression.

Another relatively obvious alternative is to use the CDF of a continuous random variable as a nonlinear regression model with a normally distributed error term. For example, Hermalin and Wallace (1994) used the normal CDF as their conditional expectation function, and Kieschnick and McCullogh (2003) used the logistic CDF. The adequacy of this approach, like the logit transformation, hinges on variance stabilization and normality of residuals. In Kieschnick and McCullogh's comparison this approach did not perform as well as beta regression. It is also the case that the bounds are not treated explicitly.

A more general—if largely atheoretical—approach is to use an automatic transformation routine, such as Tibshirani's (1988) additivity and variance stabilization (AVAS) procedure. AVAS tries to find the transformation that best linearizes the relationship between regressors and the regressand and stabilizes variance. Applying this method to our second example (in S-PLUS, 2000), we find that AVAS barely transforms anxiety while radically transforming stress. The stress transformation maps the lowest score (.01) to −0.97, while compressing scores between .05 and .37 to a tiny range from 0.866 to 0.976, whereafter the transformation describes a concave curve up to about 3.4 for a stress score of .70. The result is an exaggerated version of a logit transform that still does not entirely stabilize variance and presents rather severe problems of interpretation.

Tobit regression with censoring at both boundaries can be appropriate for modeling dependent variables that have pure boundary cases, for example, whose range is the closed interval [0, 1], whereas the applicable range for a beta-distributed variable is the open (0, 1) interval. A two-limit Tobit model (cf. Long, 1997, pp. 205–212) posits a latent variable Y that is censored at 0 and 1. For censored observations,

$$\Pr(y_i \leq 0|\mathbf{x}_i) = \Phi(-\mathbf{x}_i\beta/\sigma_i) \text{ and}$$

$$\Pr(y_i \leq 1|\mathbf{x}_i) = 1 - \Phi(1 - \mathbf{x}_i\beta/\sigma_i),$$

where Φ is the standardized normal CDF, $\mathbf{x}_i$ is a row vector of regressors from the design matrix $\mathbf{X}$, and σ_i is the standard deviation. For uncensored observations,

$$y_i = \mathbf{x}_i\beta + \varepsilon_i,$$

where ε_i is distributed $N(0, \sigma_i^2)$. An example of this kind of censoring is a scale that is the sum of a finite bank of interval-scaled items. Often the lowest possible score on that scale is not a true zero, nor is the highest possible score a true upper bound on the construct being measured.

A Tobit distribution on [0, 1] is a mixed continuous-discrete distribution, with discrete masses at 0 and 1, whereas the beta distribution is defined on (0, 1). Tobit models treat boundary cases as qualitatively distinct from cases in the interior, whereas beta regression does not. In a Tobit model, skew, heteroscedasticity, and even bimodality are accounted for by the masses at 0 and 1 rather than the diminishing returns phenomenon reflected in a beta model. Moreover, for values in (0, 1) a Tobit model will be observationally equivalent to normal regression and will not have the diminishing returns property of a beta regression. Researchers should therefore prefer a beta to a Tobit regression when the boundaries are regarded as fixed (i.e., it is not meaningful to consider out-of-domain scores and so there is no censoring) and when boundary scores are not considered as qualitatively distinct from interior scores.

From the foregoing remarks, the reading accuracy test example could be analyzed with a Tobit model with censoring at the upper end of the test scale and an appropriate dispersion submodel. It turns out that a maximum-likelihood Tobit model of these data yields much the same results as the beta regression model (details are available from Michael Smithson). We may compare these models' goodness of fit by using the BIC statistic. The results are as follows:

$$\text{Tobit BIC} = -2{*}21.602 + 7{*}\ \ln(44) = -16.715$$

$$\text{Beta BIC} = -2{*}65.902 + 7{*}\ \ln(44) = -105.315.$$

The beta regression model performs markedly better, though the Tobit model could possibly be argued for on substantive or measurement grounds. It should be noted that the Gaussian Tobit model is known to be sensitive to misspecification of the underlying distribution, though we do not at this point know how sensitive to misspecification the beta model is.

An example of an alternative conditional distribution model for a dependent variable on (0, 1) is the simplex distribution (Barndorff-Nielsen & Jorgensen, 1991). Like the beta parameterization presented here, the simplex distribution is parameterized in terms of a mean and standard deviation. Unlike the beta model, the simplex distribution is naturally a deviance-based model. Kieschnick and McCullogh (2003) compared it with their beta regression model on two data-sets and found the beta model slightly outperformed it both times in terms of goodness of fit. Of course, this approach should not be dismissed on the basis of these outcomes. When and where the simplex and beta models outperform one another is an open question. Nonetheless, it should be noted that the simplex model requires as much effort to fit and interpret as the beta model.

Cox (1996) and Papke and Wooldridge (1996) applied a quasi-likelihood approach to modeling a variable on the unit interval. The quasi-likelihood approach specifies the first and second moments of the conditional distribution as functions of the mean but does not specify the full distribution; in this sense it is a second-order analog of maximum-likelihood. This approach is useful when the relationship between the mean and variance is not captured by a standard distribution, for example, as found in Wedderburn's classic leaf blotch data, where $\text{Var}(Y) \propto \mu^2(1 - \mu)^2$ gives a much better fit than $\text{Var}(Y\ ) \propto \mu(1 - \mu)$, the variance function implied by the beta distribution (McCullagh & Nelder, 1989). Kieschnick and McCullogh (2003) compared a special case of Papke and Wooldridge's model with their beta regression model on two data sets and found that both worked reasonably well and gave similar results. However, they advised researchers to use beta regression unless the sample is large enough to justify the asymptotic arguments underlying the quasi-likelihood approaches for reasons discussed in Godfrey (1988). However, for the interested reader, we note that SAS GLIMMIX has particularly useful, very general purpose quasi-likelihood estimation facilities (SAS Institute, 2005).

It should be clear by now that we are not claiming that beta regression is always the best choice, but we have made a case that it can provide a prudent and productive alternative to current practices. It not only fits strongly skewed distributions well and handles heteroscedasticity effectively, but it also enables researchers to model both dispersion and location in a natural way. Moreover, the sets of covariates in the two submodels need not be identical, so researchers can test hypotheses about the prediction of dispersion and location separately. The beta distribution itself is a sensible alternative to the normal distribution, particularly for variables whose scales are bounded below and above. Having just two parameters, the beta distribution is as parsimonious as the normal distribution. As we noted, many psychological variables have such bounds, even if they are routinely ignored. When those bounds are meaningful or when participants' responses are affected by them, and thus exhibit diminishing returns near the boundary of the response space, beta regression should be considered. Both theoretical and empirical guidelines can aid researchers in deciding whether beta regression is appropriate or useful. Theoretical considerations include the underlying genesis of a random variable, its domain, and the nature of boundary values. Empirical considerations include several common to GLMs: goodness of fit in comparison to alternative models, sample size relative to model complexity, and the behavior of model residuals.

Starting with theoretical matters, a variable's genesis may correspond to a process that generates a beta distribution, thereby making the beta regression model a natural choice. The most familiar case is a variable of the form $Y = X_1/(X_1 + X_2)$, where X_1 and X_2 are two independent random variables with Gamma distributions. Examples of X_1 and X_2 that fit this description are amounts of time devoted to two subtasks or amounts of money allocated to two types of investment (with total task time or total investment being $X_1 + X_2$). Verkuilen (2005) discussed models of this basic form for continuous expressions of preference. Consult Johnson et al. (1995) and Gupta and Nadarajah (2004).

When a variable is bounded between 0 and 1 but its generating mechanism is unknown, the situation is more ambiguous and there is no commonly accepted distribution model. As we have suggested, a crucial distinction is between variables defined on the closed unit interval [0, 1] and those defined on the open unit interval (0, 1). For variables on (0, 1), beta regression is a viable candidate. However, variables on [0, 1] are best considered as a mixed discrete-continuous process, because observations at 0 and 1 usually are mass points. In addition to Tobit models, discrete-continuous mixture models with beta regression as the continuous component certainly are conceivable, but such models currently are unexplored.

A still more ambiguous (but very common) situation is a bounded variable with no true zero (e.g., an interval-level scale). Researchers should carefully consider the nature of the bounds, especially whether they are best regarded as fixed (i.e., whether boundary cases are censored or not) and whether the domain constitutes an open or closed interval. After all, most variables are bounded in practice, but the bounds are arbitrary and do not correspond to meaningful bounds on the underlying construct. Another practical question is how robust beta regression is under violations of continuity. Our experience thus far indicates that for responses as coarse as 7-point scales, the technique functions well, but the issue does need systematic investigation.

Turning now to empirical issues, we have made several suggestions throughout this article for evaluating the adequacy of a beta regression model and we summarize and extend them here. Starting with goodness-of-fit measures, the best guides currently are AIC and/or BIC. The log-likelihood chi-square statistic for nested model comparisons is essential in searching for the best model. Likewise, as we have done in this article, comparisons may be made between a beta regression model and any competing model whose likelihood can be computed. For individual model terms, standard errors are also a good indicator of model performance. We have found these particularly helpful in evaluating the appropriateness of dispersion submodels; unusually large standard errors suggest that the submodel may be misspecified.

The small sample behavior of beta regression currently is not well-known, so researchers will need to exercise caution in applying beta regression to small data sets. At the very least, the usual advice available regarding sample size relative to number of model parameters in GLMs should be heeded, bearing in mind that estimating location and dispersion submodels potentially introduces two parameters for every independent variable. Moreover, we recommend using both asymptotic and bootstrap estimates of parameter standard errors. Large differences between these would suggest the sample size may be too small and/or the model may be misspecified.

Finally, as with GLMs generally, model residuals provide a wealth of diagnostic information. As we have mentioned, plotting the predicted values versus the raw residuals or plotting the sorted predicted values versus the corresponding observed values is highly informative. Combining a residuals plot with predicted standard-error curves as in Figure 6 helps evaluate how well the dispersion submodel is handling variation. As mentioned earlier, we also recommend leave-one-out jackknifing or reestimation after deleting a suspect case for indications of case-wise influence on the coefficients in both submodels.

At least two extensions of beta regression seem worth exploring in the near future. One is mixed (or multilevel) and latent-variable modeling. Many commonly used models can be specified as regressions with hierarchical mixtures of distributions. Our initial forays into mixed models have indicated that estimation is more difficult than the independent observations case and may require Monte Carlo methods. The second extension is to generalizations of the beta-distribution by way of greater flexibility for models. Chapter 5 of Gupta and Nadarajah (2004) provides a useful survey of generalized beta distributions, but unfortunately none of them is readily reparameterized into a location-dispersion GLM. An alternative that we are investigating is linear mixtures of betas, which show promise in modeling multimodality as well as skew and heteroscedasticity.

References

Barndorff-Nielsen, O. E., & Jorgensen, B. (1991). Some parametric models on the simplex. *Journal of Multivariate Analysis, 39,* 106–116.

Brehm, J., & Gates, S. (1993). Donut shops and speed traps: Evaluating models of supervision on police behavior. *American Journal of Political Science, 37,* 555–581.

Buckley, J. (2002). Estimation of models with beta-distributed dependent variables: A replication and extension of Paolino (2001). *Political Analysis, 11,* 1–12.

Cleveland, W. S. (1993). *Visualizing data*. Summit, NJ: Hobart Press.

Collett, D. (2003). *Modeling binary data* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Cox, C. (1996). Nonlinear quasi-likelihood models: Applications to continuous proportions. *Computational Statistics and Data Analysis, 21,* 449–461.

Cribari-Neto, F., & Vasconcellos, K. L. P. (2002). Nearly unbiased maximum-likelihood estimation for the beta distribution. *Journal of Statistical Computation and Simulation, 72,* 107–118.

Crowder, M. J. (1978). Beta-binomial ANOVA for proportions. *Applied Statistics, 27,* 34–37.

Deady, S. (2004). *The psychological third verdict: "Not proven" or "not willing to make a decision"?* Unpublished honors thesis, The Australian National University, Canberra.

de Bustamante Simas, A. (2004). Beta regression for modeling rates and proportions (Version 1.0) [Computer software and manual]. Retrievable from http://www.cran.r-project.org

Fahrmeir, L., & Tutz, G. E. (2001). *Multivariate statistical modeling based on generalized linear models* (2nd ed.). New York: Springer.

Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research, 62,* 61–84.

Ferrari, S. L. P., & Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. *Journal of Applied Statistics, 10,* 1–18.

Godfrey, L. (1988). *Misspecifcation tests in econometrics: The Lagrange multiplier principle and other approaches.* New York: Cambridge University Press.

Greene, W. H. (2000). *Econometric analysis* (4th ed.). New York: Prentice Hall.

Gupta, A. K., & Nadarajah, S. (Eds.). (2004). *Handbook of beta distribution and its applications*. New York: Marcel Dekker.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions.* New York: Wiley.

Hardin, J. W., & Hilbe, J. W. (2003). *Generalized estimating equations*. Boca Raton, FL: Chapman & Hall/CRC.

Hermalin, B., & Wallace, N. (1994). The determinants of efficiency and solvency in savings and loans. *The RAND Journal of Economics, 25,* 361–381.

Huber, P. (1981). *Robust statistics.* New York: Wiley.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (Vol. 2, 2nd ed.). New York: Wiley.

Kieschnick, R., & McCullogh, B. D. (2003). Regression analysis of variates observed on (0,1): Percentages, proportions, and fractions. *Statistical Modeling, 3,* 193–213.

Liao, T. F. (1994). *Interpreting probability models: Logit, probit, and other generalized linear models* (Sage Series on Quantitative Applications in the Social Sciences, Vol. 101). Thousand Oaks, CA: Sage.

Long, J. S. (1997). *Regression with categorical and limited dependent variables.* Thousand Oaks, CA: Sage.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity-consistent standard errors in the linear regression model. *The American Statistician, 54,* 217–224.

Louis, W. R., Mavor, K. I., & Terry, D. J. (2003). Reflections on the statistical analysis of personality and norms in war, peace, and prejudice: Are deviant minorities the problem? *Analyses of Social Issues and Public Policy, 3,* 189–198.

Lovibond, P. F., & Lovibond, S. H. (1995). The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales with the Beck Depression and Anxiety Inventories. *Behavior Research and Therapy, 33,* 335–343.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences.* Stanford, CA: Stanford University Press.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman and Hall.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In F. Zabembka (Ed.), *Frontiers in econometrics* (pp. 105–142). New York: Academic Press.

Nagelkerke, N. J. D. (1991). A note on the general definition of the coefficient of determination. *Biometrika, 78,* 691–692.

Pammer, K., & Kevan, A. (2004). *The contribution of visual sensitivity, phonological processing and non-verbal IQ to children's reading.* Unpublished manuscript, The Australian National University, Canberra.

Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis, 9,* 325–346.

Papke, L., & Wooldridge, J. (1996). Econometric methods for fractional response variables with an application to 401(K) plan participation rates. *Journal of Applied Econometrics, 11,* 619–632.

Pham-Gia, T. (1994). Value of the beta prior information. *Communications in Statistics: Theory and Methods, 23,* 2175–2195.

Powers, D. A., & Xie, Y. (2000). *Statistical methods for categorical data.* San Diego, CA: Academic Press.

Rose, C., & Smith, M. D. (2002). *Mathematical statistics with mathematica.* New York: Springer.

SAS Institute. (2005). *The GLIMMIX procedure.* Cary, NC: Author.

Schönemann, P. H. (1983). Some theory and results for metrics for bounded response scales. *Journal of Mathematical Psychology, 27,* 311–324.

Smithson, M. (2005). Fuzzy set inclusion: Linking fuzzy set methods with mainstream techniques. *Sociological Methods and Research, 33,* 431–461.

Smithson, M., & Verkuilen, J. (2006). *Fuzzy set theory: Applications in the social sciences* (Sage Series on Quantitative Applications in the Social Sciences, Vol. 147). Thousand Oaks, CA: Sage.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society, Series B, 51,* 47–60.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis.* Thousand Oaks, CA: Sage.

Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing.* New York: Wiley-Interscience.

Tibshirani, R. (1988). Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association, 83,* 394–405.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison Wesley.

Verkuilen, J. (2005, August). *The Beta-BTL: A continuous analog to the Bradley-Terry-Luce model*. Paper presented at the annual meeting of the Society for Mathematical Psychology, Memphis, TN.

Weisstein, E. W. (2005). *Mathworld*. Retrieved September 1, 2005, from http://www.mathworld.wolfram.com/BetaDistribution.html

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53,* 300–314.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego, CA: Academic Press.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, NJ: Erlbaum.

Wolfram Research. (2005). Gamma, beta, erf (3,906 formulas). Retrieved September 1, 2005, from http://functions.wolfram.com/GammaBetaErf/

Received May 14, 2005
Revision received November 18, 2005
Accepted November 18, 2005 ■