Journal of Applied Statistics,
Vol. 31, No. 7, 799–815, August 2004

Routledge
Taylor & Francis Group

Beta Regression for Modelling Rates and Proportions

SILVIA L. P. FERRARI* AND FRANCISCO CRIBARI-NETO**
**Departamento de Estatística/IME, Universidade de São Paulo, Brazil, **Departamento de Estatística, CCEN, Universidade Federal de Pernambuco, Brazil*

ABSTRACT *This paper proposes a regression model where the response is beta distributed using a parameterization of the beta law that is indexed by mean and dispersion parameters. The proposed model is useful for situations where the variable of interest is continuous and restricted to the interval (0, 1) and is related to other variables through a regression structure. The regression parameters of the beta regression model are interpretable in terms of the mean of the response and, when the logit link is used, of an odds ratio, unlike the parameters of a linear regression that employs a transformed response. Estimation is performed by maximum likelihood. We provide closed-form expressions for the score function, for Fisher's information matrix and its inverse. Hypothesis testing is performed using approximations obtained from the asymptotic normality of the maximum likelihood estimator. Some diagnostic measures are introduced. Finally, practical applications that employ real data are presented and discussed.*

KEY WORDS: Beta distribution, maximum likelihood estimation, leverage, proportions, residuals

Introduction

Practitioners commonly use regression models to analyse data that are perceived to be related to other variables. The linear regression model, in particular, is commonly used in applications. It is not, however, appropriate for situations where the response is restricted to the interval (0, 1) since it may yield fitted values for the variable of interest that exceed its lower and upper bounds. A possible solution is to transform the dependent variable so that it assumes values on the real line, and then to model the mean of the transformed response as a linear predictor based on a set of exogenous variables. This approach, however, has drawbacks, one of them being the fact that the model parameters cannot be easily interpreted in terms of the original response. Another shortcoming is that measures of proportions typically display asymmetry, and hence inference based on the normality assumption can be misleading. Our goal is to propose a regression model that is tailored for situations where the dependent variable (y) is measured continuously on the standard unit interval, i.e. $0<y<1$. The proposed model is based on the assumption that the response is beta distributed. The beta distribution, as is well known, is very flexible for modelling proportions

Correspondence Address: Silvia L. P. Ferrari, Departamento de Estatística/IME, Universidade de São Paulo, Caixa Postal 66281, São Paulo/SP, 05311-970, Brazil. Email: sferrari@ime.usp.br

0266-4763 Print/ 1360-0532 Online/04/070799-17
DOI: 10.1080/0266476042000214501

since its density can have quite different shapes depending on the values of the two parameters that index the distribution. The beta density is given by

$$\pi(y;p,q)=\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}y^{p-1}(1-y)^{q-1}, \quad 0<y<1 \tag{1}$$

where $p>0$, $q>0$ and $\Gamma(\cdot)$ is the gamma function. The mean and variance of y are, respectively,

$$E(y)=\frac{p}{(p+q)} \tag{2}$$

and

$$\mathrm{var}(y)=\frac{pq}{(p+q)^2(p+q+1)} \tag{3}$$

The mode of the distribution exists when both p and q are greater than one: $\mathrm{mode}(y)=(p-1)/(p+q-2)$. The uniform distribution is a particular case of equation (1) when $p=q=1$. Estimation of p and q by maximum likelihood and the application of small sample bias adjustments to the maximum likelihood estimators of these parameters are discussed by Cribari-Neto & Vasconcellos (2002).

'Beta distributions are very versatile and a variety of uncertainties can be usefully modelled by them. This flexibility encourages its empirical use in a wide range of applications' (Johnson *et al.*, 1995, p. 235). Several applications of the beta distribution are discussed by Bury (1999) and by Johnson *et al.* (1995). These applications, however, do not involve situations where the practitioner is required to impose a regression structure for the variable of interest. Our interest lies in situations where the behaviour of the response can be modelled as a function of a set of exogenous variables. To that end, we shall propose a beta regression model. We shall also discuss the estimation of the unknown parameters by maximum likelihood and some diagnostic techniques. Large sample inference is also considered. The modelling and inferential procedures we propose are similar to those for generalized linear models (McCullagh & Nelder, 1989), except that the distribution of the response is not a member of the exponential family. An alternative to the model we propose is the simplex model in Jørgensen (1997) which is defined by four parameters. Our model, on the other hand, is defined by only two parameters, and is flexible enough to handle a wide range of applications.

It is noteworthy that several empirical applications can be handled using the proposed class of regression models. As a first illustration, consider the dataset collected by Prater (1956). The dependent variable is the proportion of crude oil converted to gasoline after distillation and fractionation, and the potential covariates are: the crude oil gravity (degrees API), the vapour pressure of the crude oil ($\mathrm{lbf/in^2}$), the crude oil 10% point ASTIM (i.e. the temperature at which 10% of the crude oil has become vapour), and the temperature (°F) at which all the gasoline is vaporized. The dataset contains 32 observations on the response and on the independent variables. It has been noted (Daniel & Wood, 1971, Ch. 8)

that there are only ten sets of values of the first three explanatory variables that correspond to ten different crudes and were subjected to experimentally controlled distillation conditions. This dataset was analysed by Atkinson (1985), who used the linear regression model and noted that there is 'indication that the error distribution is not quite symmetrical, giving rise to some unduly large and small residuals' (Atkinson, 1995, p. 60). He proceeded to transform the response so that the transformed dependent variable assumed values on the real line, and then used it in a linear regression analysis. Our approach will be different: we shall analyse these data using the beta regression model proposed in the next section.

The paper unfolds as follows. The next section presents the beta regression model, and discusses maximum likelihood estimation and large sample inference. Diagnostic measures are discussed in the section after. The fourth section contains applications of the proposed regression model, including an analysis of Prater's gasoline data. Concluding remarks are given in the final section. Technical details are presented in two separate appendices.

The Model, Estimation and Testing

Our goal is to define a regression model for beta distributed random variables. The density of the beta distribution is given in equation (1), where it is indexed by p and q. However, the regression it is typically more useful to model the mean of the response. It is also typical to define the model so that it contains a precision (or dispersion) parameter. In order to obtain a regression structure for the mean of the response along with a precision parameter, we shall work with a different parameterization of the beta density. Let $\mu = p/(p+q)$ and $\phi = p+q$, i.e. $p = \mu\phi$ and $q = (1-\mu)\phi$. It follows from equations (2) and (3) that

$$\mathrm{E}(y) = \mu$$

and

$$\mathrm{var}(y) = \frac{V(\mu)}{1+\phi}$$

where $V(\mu) = \mu(1-\mu)$, so that μ is the mean of the response variable and ϕ can be interpreted as a precision parameter in the sense that, for fixed μ, the larger the value of ϕ, the smaller the variance of y. The density of y can be written, in the new parameterization, as

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0<y<1 \qquad (4)$$

where $0<\mu<1$ and $\phi>0$. Figure 1 shows a few different beta densities along with the corresponding values of (μ, ϕ). It is noteworthy that the densities can display quite different shapes depending on the values of the two parameters. In particular, it can be symmetric (when $\mu = 1/2$) or asymmetric (when $\mu \neq 1/2$). Additionally, we note that the dispersion of the distribution, for fixed μ, decreases as ϕ increases. It is also interesting to note that, in the two upper panels, two

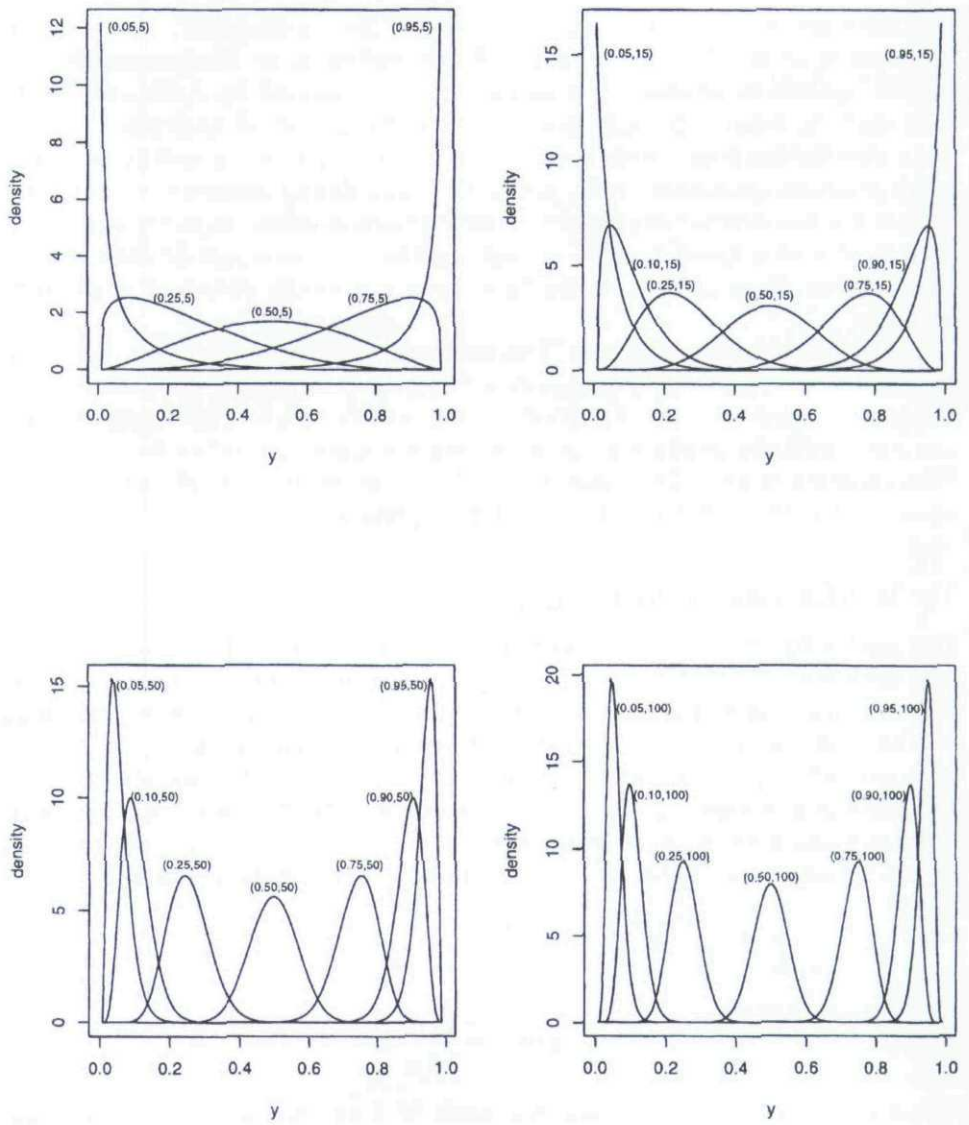

Figure 1. Beta densities for different combinations of (μ, ϕ)

densities have 'J shapes' and two others have inverted 'J shapes'. Although we did not plot the uniform case, we note that when $\mu = 1/2$ and $\phi = 2$ the density reduces to that of a standard uniform distribution. The beta density can also be 'U shaped' (skewed or not), and this situation is also not displayed in Figure 1.

Throughout the paper we shall assume that the response is constrained to the standard unit interval (0, 1). The model we shall propose, however, is still useful for situations where the response is restricted to the interval (a, b), where a and b are known scalars, $a < b$. In this case, one would model $(y-a)/(b-a)$ instead of modelling y directly.

Let $y_1, \ldots, y_n$ be independent random variables, where each y_t, $t=1, \ldots, n$, follows the density in equation (4) with mean μ_t and unknown precision ϕ. The model is obtained by assuming that the mean of y_t can be written as

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_t \tag{5}$$

where $\beta = (\beta_1, \ldots, \beta_k)^\top$ is a vector of unknown regression parameters ($\beta \in \mathbb{R}k$) and $x_{t1}, \ldots, x_{tk}$ are observations on k covariates ($k < n$), which are assumed fixed and known. Finally, $g(\cdot)$ is a strictly monotonic and twice differentiable link function that maps (0, 1) into $\mathbb{R}$. Note that the variance of y_t is a function of μ_t and, as a consequence, of the covariate values. Hence, non-constant response variances are naturally accommodated into the model.

There are several possible choices for the link function $g(\cdot)$. For instance, one can use the logit specification $g(\mu) = \log\{\mu/(1-\mu)\}$, the probit function $g(\mu) = \Phi^{-1}(\mu)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable, the complementary log-log link $g(\mu) = \log\{-\log(1-\mu)\}$, the log-log link $g(\mu) = -\log\{-\log(\mu)\}$, among others. For a comparison of these link functions, see McCullagh & Nelder (1989, section 4.3.1), and for other transformations, see Atkinson (1985, Ch. 7).

A particularly useful link function is the logit link, in which case we case write

$$\mu_t = \frac{e^{x_t^\top \beta}}{1 + e^{x_t^\top \beta}}$$

where $x_t^\top = (x_{t1}, \ldots, x_{tk})$, $t = 1, \ldots, n$. Here, the regression parameters have an important interpretation. Suppose that the value of the ith regressor is increased by c units and all other independent variables remain unchanged, and let $\mu\dagger$ denote the mean of y under the new covariate values, whereas μ denotes the mean of y under the original covariate values. Then, it is easy to show that

$$e^{c\beta_i} = \frac{\mu\dagger/(1-\mu\dagger)}{\mu/(1-\mu)}$$

that is, $\exp\{c\beta_i\}$ equals the odds ratio. Consider, for instance, Prater's gasoline example introduced in the previous section, and define the odds of converting crude oil into gasoline as the number of units of crude oil, out of ten units, that are, on average, converted into gasoline divided by the number of units that are not converted. As an illustration, if, on average, 20% of the crude oil is transformed into gasoline, then the odds of conversion equals 2/8. Suppose that the temperature at which all the gasoline is vaporized increases by 50°F, then 50 times the regression parameter associated with this covariate can be interpreted as the log of the ratio between the chance of converting crude oil into gasoline under the new setting relative to the old setting, all other variables remaining constant.

The log-likelihood function based on a sample of n independent observations is

$$\ell(\beta, \phi) = \sum_{t=1}^{n} \ell_t(\mu_t, \phi) \tag{6}$$

where

$$\ell_t(\mu_t, \phi) = \log\Gamma(\phi) - \log\Gamma(\mu_t\phi) - \log\Gamma((1-\mu_t)\phi) + (\mu_t\phi - 1)\log y_t \tag{7}$$
$$+ \{(1-\mu_t)\phi - 1\}\log(1-y_t)$$

with μ_t defined so that equation (5) holds. Let $y_t^* = \log\{y_t/(1-y_t)\}$ and $\mu_t^* = \psi(\mu_t\phi) - \psi((1-\mu_t)\phi)$. The score function, obtained by differentiating the log-likelihood function with respect to the unknown parameters (see Appendix A), is given by $(U_\beta(\beta, \phi)^\top, U_\phi(\beta, \phi))^\top$, where

$$U_\beta(\beta, \phi) = \phi X^\top T(y^* - \mu^*) \tag{8}$$

with X being an $n \times k$ matrix whose tth row is $x_t^\top$, $T = \text{diag}\{1/g'(\mu_1), \ldots, 1/g'(\mu_n)\}$, $y^* = (y_1^*, \ldots, y_n^*)^\top$ and $\mu^* = (\mu_1^*, \ldots, \mu_n^*)^\top$, and

$$U_\phi(\beta, \phi) = \sum_{t=1}^{n} \{\mu_t(y_t^* - \mu_t^*) + \log(1-y_t) - \psi((1-\mu_t)\phi) + \psi(\phi)\} \tag{9}$$

The next step is to obtain an expression for Fisher's information matrix. The notation can be described as follows. Let $W = \text{diag}\{w_1, \ldots, w_n\}$, with

$$w_t = \phi\{\psi'(\mu_t\phi) + \psi'((1-\mu_t)\phi)\}\frac{1}{\{g'(\mu_t)\}^2}$$

$c = (c_1, \ldots, c_n)^\top$, with $c_t = \phi\{\psi'(\mu_t\phi)\mu_t - \psi'((1-\mu_t)\phi)(1-\mu_t)\}$, where $\psi'(\cdot)$ is the trigamma function. Also, let $D = \text{diag}\{d_1, \ldots, d_n\}$, with $d_t = \psi'(\mu_t\phi)\mu_t^2 + \psi'((1-\mu_t)\phi)(1-\mu_t)^2 - \psi'(\phi)$. It is shown in Appendix A that Fisher's information matrix is given by

$$K = K(\beta, \phi) = \begin{pmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{pmatrix} \tag{10}$$

where $K_{\beta\beta} = \phi X^\top W X$, $K_{\beta\phi} = K_{\phi\beta}^\top = X^\top T c$ and $K_{\phi\phi} = \text{tr}(D)$. Note that the parameters β and ϕ are not orthogonal, in contrast to what is verified in the class of generalized linear regression models (McCullagh & Nelder, 1989).

Under the usual regularity conditions for maximum likelihood estimation, when the sample size is large,

$$\begin{pmatrix} \hat\beta \\ \hat\phi \end{pmatrix} \underset{\sim}{\mathcal{N}}_{k+1}\left(\begin{pmatrix} \beta \\ \phi \end{pmatrix}, K^{-1}\right)$$

approximately, where $\hat\beta$ and $\hat\phi$ are the maximum likelihood estimators of β and ϕ, respectively. It is thus useful to obtain an expression for K^{-1}, which can be

used to obtain asymptotic standard errors for the maximum likelihood estimates. Using standard expressions for the inverse of partitioned matrices (e.g. Rao, 1973, p. 33), we obtain

$$K^{-1} = K^{-1}(\beta, \phi) = \begin{pmatrix} K^{\beta\beta} & K^{\beta\phi} \\ K^{\phi\beta} & K^{\phi\phi} \end{pmatrix} \tag{11}$$

where

$$K^{\beta\beta} = \frac{1}{\phi}(X^\top WX)^{-1}\left\{I_k + \frac{X^\top Tcc^\top T^\top X(X^\top WX)^{-1}}{\gamma\phi}\right\}$$

with $\gamma = \mathrm{tr}(D) - \phi^{-1}c^\top T^\top X(X^\top WX)^{-1}X^\top Tc$,

$$K^{\beta\phi} = (K^{\phi\beta})^\top = -\frac{1}{\gamma\phi}(X^\top WX)^{-1}X^\top Tc$$

and $K^{\phi\phi} = \gamma^{-1}$. Here, I_k is the $k \times k$ identity matrix.

The maximum likelihood estimators of β and ϕ are obtained from the equations $U_\beta(\beta, \phi) = 0$ and $U_\phi(\beta, \phi) = 0$, and do not have closed-form. Hence, they need to be obtained by numerically maximizing the log-likelihood function using a nonlinear optimization algorithm, such as a Newton algorithm or a quasi-Newton algorithm; for details, see Nocedal & Wright (1999). The optimization algorithms require the specification of initial values to be used in the iterative scheme. Our suggestion is to use as an initial point estimate for β, the ordinary least squares estimate of this parameter vector obtained from a linear regression of the transformed responses $g(y_1), \ldots, g(y_n)$ on X, i.e. $(X^\top X)^{-1}X^\top z$, where $z = (g(y_1), \ldots, g(y_n))^\top$. We also need an initial guess for ϕ. As noted earlier, $\mathrm{var}(y_t) = \mu_t(1-\mu_t)/(1+\phi)$ which implies that $\phi = \mu_t(1-\mu_t)/\mathrm{var}(y_t) - 1$. Note that

$$\mathrm{var}(g(y_t)) \approx \mathrm{var}\{g(\mu_t) + (y_t - \mu_t)g'(\mu_t)\} = \mathrm{var}(y_t)\{g'(\mu_t)\}^2$$

that is, $\mathrm{var}(y_t) \approx \mathrm{var}\{g(y_t)\}\{g'(\mu_t)\}^{-2}$. Hence, the initial guess for ϕ we suggest is

$$\frac{1}{n}\sum_{t=1}^{n}\frac{\check{\mu}_t(1-\check{\mu}_t)}{\check{\sigma}_t^2} - 1$$

where $\check{\mu}_t$ is obtained by applying $g^{-1}(\cdot)$ to the tth fitted value from the linear regression of $g(y_1), \ldots, g(y_n)$ on X, i.e. $\check{\mu}_t = g^{-1}(x_t^\top(X^\top X)^{-1}X^\top z)$ and $\check{\sigma}_t^2 = \check{e}^\top\check{e}/[(n-k)\{g'(\check{\mu}_t)\}^2]$; here, $\check{e} = z - X(X^\top X)^{-1}X^\top z$ is the vector of ordinary least squares residuals from the linear regression that employs the transformed response. These initial guesses worked well in the applications described in the fourth section.

Large sample inference is considered in Appendix B. We have developed likelihood ratio, score and Wald tests for the regression parameters. In addition, we have obtained confidence intervals for the precision and for the regression parameters, for the odds ratio when the logit link is used, and for the mean response.

Diagnostic Measures

After the fit of the model, it is important to perform diagnostic analyses in order to check the goodness-of-fit of the estimated model. We shall introduce a global measure of explained variation and graphical tools for detecting departures from the postulated model and influential observations.

At the outset, a global measure of explained variation can be obtained by computing the pseudo R^2 (R_p^2) defined as the square of the sample correlation coefficient between $\hat{\eta}$ and $g(y)$. Note that $0 \leqslant R_p^2 \leqslant 1$ and perfect agreement between $\hat{\eta}$ and $g(y)$, and hence between $\hat{\mu}$ and y, yields $R_p^2 = 1$.

The discrepancy of a fit can be measured as twice the difference between the maximum log-likelihood achievable (saturated model) and that achieved by the model under investigation. Let $D(y; \mu, \phi) = \Sigma_{t=1}^{n} 2(\ell_t(\tilde{\mu}_t, \phi) - \ell_t(\mu_t, \phi))$, where $\tilde{\mu}_t$ is the value of μ_t that solves $\partial \ell_t / \partial \mu_t = 0$, i.e. $\phi(y_t^* - \mu_t^*) = 0$. When ϕ is large, $\mu_t^* \approx \log\{\mu_t/(1-\mu_t)\}$, and it then follows that $\tilde{\mu}_t \approx y_t$; see Appendix B. For known ϕ, this discrepancy measure is $D(y; \bar{\mu}, \phi)$, where $\bar{\mu}$ is the maximum likelihood estimator of μ under the model being investigated. When ϕ is unknown, an approximation to this quantity is $D(y; \hat{\mu}, \hat{\phi})$; it can be named, as usual, the deviance for the current mode. Note that $D(y; \hat{\mu}, \hat{\phi}) = \Sigma_{t=1}^{n} (r_t^d)^2$, where

$$r_t^d = \text{sign}(y_t - \hat{\mu}_t)\{2(\ell_t(\tilde{\mu}_t, \hat{\phi}) - \ell_t(\hat{\mu}_t, \hat{\phi}))\}^{1/2}$$

Note now that the tth observation contributes a quantity $(r_t^d)^2$ to the deviance, and thus an observation with a large absolute value of r_t^d can be viewed as discrepant. We shall call r_t^d the tth deviance residual.

It is also possible to define the standardized residuals:

$$r_t = \frac{y_t - \hat{\mu}_t}{\sqrt{\hat{\text{var}}(y_t)}}$$

where $\hat{\mu}_t = g^{-1}(x_t^{\top}\hat{\beta})$ and $\hat{\text{var}}(y_t) = \{\hat{\mu}_t(1-\hat{\mu}_t)\}/(1+\hat{\phi})$. A plot of these residuals against the index of the observations (t) should show no detectable pattern. Also, a detectable trend in the plot of r_t against $\hat{\eta}_t$ could be suggestive of link function misspecification.

Since the distribution of the residuals is not known, half-normal plots with simulated envelopes are a helpful diagnostic tool (Atkinson, 1985, section 4.2; Neter *et al.*, 1996, section 14.6). The main idea is to enhance the usual half-normal plot by adding a simulated envelope that can be used to decide whether the observed residuals are consistent with the fitted model. Half-normal plots with a simulated envelope can be produced as follows:

(i) fit the model and generate a simulated sample of n independent observations using the fitted model as if it were the true model;
(ii) fit the model to the generated sample, and compute the ordered absolute values of the residuals;
(iii) repeat steps (i) ad (ii) k times;
(iv) consider the n sets of the k order statistics; for each set compute its average, minimum and maximum values;

(v) plot these values and the ordered residuals of the original sample against the half-normal scores $\Phi^{-1}((t+n-1/8)/(2n+1/2))$.

The minimum and maximum values of the k order statistics yield the envelope. Atkinson (1985, p. 36) suggests using $k=19$, so that the probability that a given absolute residual will fall beyond the upper band provided by the envelope is approximately equal to $1/20=0.05$. Observations corresponding to absolute residuals outside the limits provided by the simulated envelope are worthy of further investigation. Additionally, if a considerable proportion of points falls outside the envelope, then one has evidence against the adequacy of the fitted model.

Next, we shall be concerned with the identification of influential observations and residual analysis. In what follows we shall use the generalized leverage proposed by Wei *et al.* (1998), which is defined as

$$\mathrm{GL}(\tilde{\theta})=\frac{\partial \tilde{y}}{\partial y^{\top}}$$

where θ is an s-vector such that $\mathrm{E}(y)=\mu(\theta)$ and $\tilde{\theta}$ is an estimator of θ, with $\tilde{y}=\mu(\tilde{\theta})$. Here, the (t, u) element of $\mathrm{GL}(\tilde{\theta})$, i.e. the generalized leverage of the estimator $\tilde{\theta}$ at (t, u), is the instantaneous rate of change in tth predicted value with respect to the uth response value. As noted by the authors, the generalized leverage is invariant under reparameterization and observations with large GL_{tu} are leverage points. Let $\hat{\theta}$ be the maximum likelihood estimator of θ, assumed to exist and to be unique, and assume that the log-likelihood function has second-order continuous derivatives with respect to θ and y. Wei *et al.* (1998) have shown that the generalized leverage is obtained by evaluating

$$\mathrm{GL}(\theta)=D_{\theta}\left(-\frac{\partial^2 \ell}{\partial\theta\partial\theta^{\top}}\right)^{-1}\frac{\partial^2 \ell}{\partial\theta\partial y^{\top}}$$

at $\hat{\theta}$, where $D_{\theta}=\partial\mu/\partial\theta^{\top}$.

As a first step, we shall obtain a closed-form for $\mathrm{GL}(\beta)$ in the beta regression model proposed in the previous section under the assumption that ϕ is known. It is easy to show that $D_{\beta}=TX$. The expression for the elements of $-\partial^2\ell/\partial\beta\partial\beta^{\top}$ is given in Appendix A, and it follows that

$$-\frac{\partial^2 \ell}{\partial\beta\partial\beta^{\top}}=\phi X^{\top}QX$$

where $Q=\mathrm{diag}\{q_1,\ldots,q_n\}$ with

$$q_t=\left[\phi\{\psi'(\mu_t\phi)+\psi'((1-\mu_t)\phi\}+(y_t^*-\mu_t^*)\frac{g''(\mu_t)}{g'(\mu_t)}\right]\frac{1}{\{g'(\mu_t)\}^2},\quad t=1,\ldots,n$$

Additionally, it can be shown that $\partial^2\ell/\partial\beta\partial y^{\top}=\phi X^{\top}TM$, where $M=\mathrm{diag}\{m_1,\ldots,m_n\}$ with $m_t=1/\{y_t(1-y_t)\}$, $t=1,\ldots,n$. Therefore, we obtain

$$\mathrm{GL}(\beta)=TX(X^{\top}QX)^{-1}X^{\top}TM \tag{12}$$

We note that if we replace the observed information, $-\partial^2\ell/\partial\beta\partial\beta^\top$, by the expected information, $\mathrm{E}(-\partial^2\ell/\partial\beta\partial\beta^\top)$, the expression for $\mathrm{GL}(\beta)$ is as given in equation (12) but with Q replaced by W; we shall call this matrix $\mathrm{GL}^*(\beta)$. It is noteworthy that the diagonal elements of $\mathrm{GL}^*(\beta)$ are the same as those of $M^{1/2}TX(X^\top WX)^{-1}X^\top M^{1/2}$, and that $M^{1/2}T$ is a diagonal matrix whose tth diagonal element is given by $\{g'(\mu_t)V(y_t)^{1/2}\}^{-1}$. It is important to note that there is a close connection between the diagonal elements of $\mathrm{GL}^*(\beta)$ and those of the usual 'hat matrix',

$$H = W^{1/2}X(X^\top WX)^{-1}X^\top W^{1/2}$$

when ϕ is large. The relationship stems from the fact that, when the precision parameter is large, the tth diagonal elements of $W^{1/2}$ is approximately equal to $\{g'(\mu_t)V(\mu_t)^{1/2}\}^{-1}$; see Appendix C.

Now let ϕ be unknown, and hence $\theta^\top = (\beta^\top, \phi)$. Here, $D_\theta = [TX\ 0]$, where 0 is an n-vector of zeros. Also, $-\partial^2\ell/\partial\theta\partial\theta^\top$ is given by equation (10) with W replaced by Q and c replaced by f, where $f = (f_1, \ldots, f_n)^\top$ with $f_t = \{c_t - (y_t^* - \mu_t^*)\}$, $t = 1, \ldots, n$. It is thus clear that the inverse of $-\partial^2\ell/\partial\theta\partial\theta^\top$ will be given by equation (11) with W and c replaced by Q and f, respectively. Additionally,

$$\frac{\partial^2\ell}{\partial\theta\partial y^\top} = \begin{pmatrix} \phi X^\top TM \\ b^\top \end{pmatrix}$$

where $b = (b_1, \ldots, b_n)^\top$ with $b_t = -(y_t - \mu_t)/\{y_t(1-y_t)\}$, $t = 1, \ldots, n$. It can now be shown that

$$\mathrm{GL}(\beta, \phi) = \mathrm{GL}(\beta) + \frac{1}{\gamma\phi} TX(X^\top QX)^{-1}X^\top Tf(f^\top TX(X^\top QX)^{-1}X^\top TM - b^\top)$$

where $\mathrm{GL}(\beta)$ is given in equation (12). When ϕ is large, $\mathrm{GL}(\beta, \phi) \approx \mathrm{GL}(\beta)$.

A measure of the influence of each observation on the regression parameter estimates is Cook's distance (Cook, 1977) given by $k^{-1}(\hat\beta - \hat\beta_{(t)})^\top X^\top WX(\hat\beta - \hat\beta_{(t)})$, where $\hat\beta_{(t)}$ is the parameter estimate without the tth observation. It measures the squared distance between $\hat\beta$ and $\hat\beta_{(t)}$. To avoid fitting the model $n+1$ times, we shall use the usual approximation to Cook's distance given by

$$C_t = \frac{h_{tt}r_t^2}{k(1-h_{tt})^2}$$

It combines leverage and residuals. It is common practice to plot C_t against t.

Finally, we note that other diagnostic measures can be considered, such as local influence measures (Cook, 1986).

Applications

This section contains two applications of the beta regression model proposed in the second section. all computations were carried out using the matrix programming language Ox (Doornik, 2001). The computer code and dataset used in the first application are available at http://www.de.ufpe.br/~cribari/betareg_example.zip.

Table 1. Parameter estimates using Prater's gasoline data

Parameter	Estimate	Std. error	z stat	p-value
β_1	−6.15957	0.18232	−33.78	0.0000
β_2	1.72773	0.10123	17.07	0.0000
β_3	1.32260	0.11790	11.22	0.0000
β_4	1.57231	0.11610	13.54	0.0000
β_5	1.05971	0.10236	10.35	0.0000
β_6	1.13375	0.10352	10.95	0.0000
β_7	1.04016	0.10604	9.81	0.0000
β_8	0.54369	0.10913	4.98	0.0000
β_9	0.49590	0.10893	4.55	0.0000
β_{10}	0.38579	0.11859	3.25	0.0011
β_{11}	0.01097	0.00041	26.58	0.0000
ϕ	440.27838	110.02562		

Estimation was performed using the quasi-Newton optimization algorithm known as BFGS with analytic first derivatives. The choice of starting values for the unknown parameters followed the suggestion made in the second section.

Consider initially Prater's gasoline data described in the Introduction. The interest lies in modelling the proportion of crude oil converted to gasoline after distillation and fractionation. As noted earlier, there are only ten sets of values of three of the explanatory variables which correspond to ten different crudes subjected to experimentally controlled distillation conditions. The data were ordered according to the ascending order of the covariate that measures the temperature at which 10% of the crude oil has become vapour. This variable assumes ten different values and they are used to define the ten batches of crude oil. The model specification for the mean of the response uses an intercept ($x_1 = 1$), nine dummy variables for the first nine batches of crude oil ($x_2, \ldots, x_{10}$) and the covariate that measures the temperature (°F) at which all the gasoline is vaporized (x_{11}). Estimation results using the logit link function are given in Table 1.

The pseudo R^2 of the estimated regression was 0.9617. Diagnostic plots are given in Figure 2. An inspection of Figure 2 reveals that the largest standardized and deviance residuals in absolute value correspond to observation 4. Also, C_4 is much larger than the remaining Cook's measures, thus suggesting that the fourth observation is the most influential. Additionally, observation 4 deviates from the pattern shown in the lower right panel (plot of the diagonal elements of $\mathrm{GL}(\hat{\beta}, \hat{\phi})$ *v.* $\hat{\mu}_t$; observation 29, the one with largest generalized leverage, also displays deviation from the main pattern). On the other hand, it is noteworthy that the generalized leverage for this observation is not large relative to the remaining ones. We note, however, that y_4 is the largest value of the response; its observed value is 0.457 and the corresponding fitted value equals 0.508. The analysis of these data carried out by Atkinson (1985, Ch. 7) using a linear regression specification for transformations of the response also singles out observation 4 as influential.

We fitted the beta regression model without the fourth observation and noted

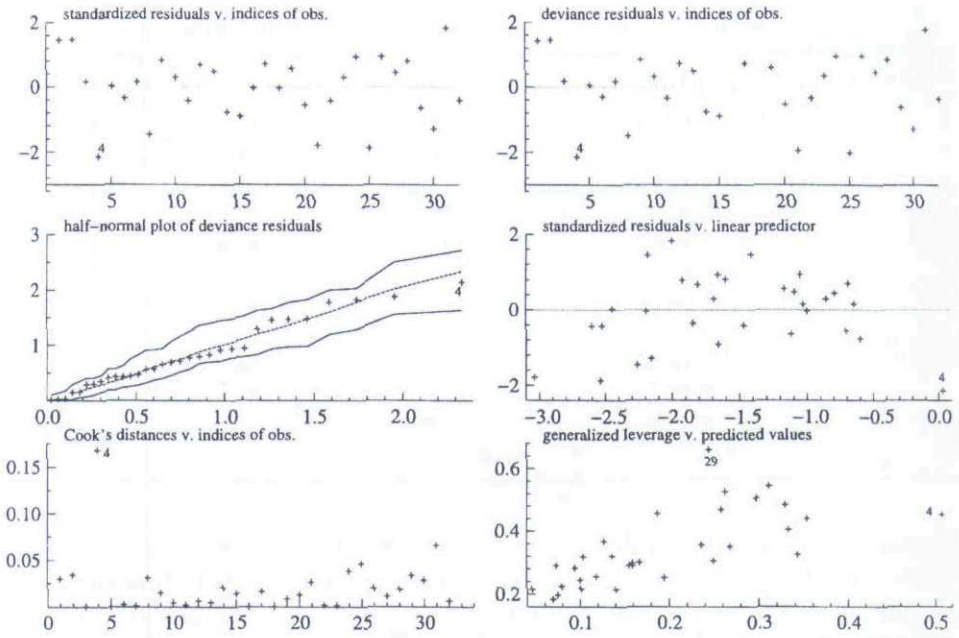

Figure 2. Six diagnostic plots for Prater's gasoline data. The upper left panel plots the standardized residuals against t, the upper right panel plots the deviance residuals versus t, the middle left panel displays the half-normal plot of absolute deviance residuals with a simulated envelope, the middle right panel plots standardized residuals against $\hat{\eta}_t$, the lower left panel presents a plot of C_t versus t, and the lower right panel plots the diagonal elements of $\mathrm{GL}(\hat{\beta}, \phi)$ against $\hat{\mu}_t$

that the point estimates of the βs were not significantly altered, but that the estimate of the precision parameter jumped from 440.3 to 577.8; despite that, however, the reduction in the asymptotic standard errors of the regression parameter estimates was negligible.

The next application uses data on food expenditure, income, and number of persons in each household from a random sample of 38 households in a large US city; the source of the data is Griffiths *et al.* (1993, Table 15.4). The interest lies in modelling the proportion of income spent on food (y) as a function of the level of income (x_2) and the number of persons in the household (x_3). At the outset, consider a linear regression of the response on the covariates. The estimated regression displayed evidence of heteroskedasticity; the p-value for Koenker's (1981) homoskedasticity test was 0.0514. If we consider instead the regression of $\log\{y/(1-y)\}$ on the two covariates, the evidence of heteroskedasticity is attenuated, but the residuals become highly asymmetric to the left.

We shall now consider the beta regression model proposed in the second section. As previously mentioned, this model accommodates naturally non-constant variances and skewness. The model is specified as

$$g(\mu_t) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3}$$

Table 2. Parameter estimates using data on food expenditure

Parameter	Estimate	Std. error	z stat	p-value
β_1	−0.62255	0.22385	−2.78	0.0054
β_2	−0.01230	0.00304	−4.05	0.0001
β_3	0.11846	0.03534	3.35	0.0008
ϕ	35.60975	8.07960		

The link function used was logit. The parameter estimates are given in Table 2. The pseudo R^2 of the estimated regression was 0.3878.

The values in Table 2 show that both covariates are statistically significant at the usual nominal levels. We also note that there is a negative relationship between the mean response (proportion of income spent on food) and the level of income, and that there is a positive relationship between the mean response and the number of persons in the household. Diagnostic plots similar to those presented in Figure 2 were also produced but for brevity are not presented.

Concluding Remarks

This paper proposed a regression model tailored for responses that are measured continuously on the standard unit interval, i.e. $y \in (0, 1)$, which is the situation that practitioners encounter when modelling rates and proportions. The underlying assumption is that the response follows a beta law. As is well known, the beta distribution is very flexible for modelling data on the standard unit interval, since the beta density can display quite different shapes depending on the values of the parameters that index the distribution. We use a parameterization in which a function of the mean of the dependent variable is given by a linear predictor that is defined by regression parameters and explanatory variables. The proposed parameterization also allows for a precision parameter. When the logit link function is used to transform the mean response, the regression parameters can be interpreted in terms of the odds ratio. Parameter estimation is performed by maximum likelihood, and we provide closed-form expressions for the score function, for Fisher's information matrix and its inverse. Interval estimation for different population quantities (such as regression parameters, precision parameter, mean response, odds ratio) is discussed. Tests of hypotheses on the regression parameters can be performed using asymptotic tests, and three tests are presented: likelihood ratio, score and Wald. We also consider a set of diagnostic techniques that can be employed to identify departures from the postulated model and influential observations. These include a measure of the degree of leverage of the different observations, and a half normal plot of residuals with envelopes obtained from a simulation scheme. Applications using real data sets were presented and discussed.

Acknowledgements

The authors gratefully acknowledge partial financial support from CNPq and FAPESP. The authors also thank Gilberto Paula and a referee for comments and suggestions on an earlier draft.

References

Abramowtiz, M. & Stegun, I. A. (1965) *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables* (New York: Dover).

Atkinson, A. C. (1985) *Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis* (New York: Oxford University Press).

Bury, K. (1999) *Statistical Distributions in Engineering* (New York: Cambridge University Press).

Cook, R. D. (1977) Detection of influential observations in linear regression, *Technometrics*, 19, pp. 15–18.

Cook, R. D. (1986) Assessment of local influence (with discussion), *Journal of the Royal Statistical Society B*, 48, pp. 133–169.

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

Daniel, C. & Wood, F. S. (1971) *Fitting Equations to Data* (New York: Wiley).

Doornik, J. A. (2001) *Ox: an Object-oriented Matrix programming Language*, 4th edn (London: Timberlake Consultants and Oxford: http://www.nuff.ox.ac.uk/Users/Doornik/).

Griffiths, W. E., Hill, R. C. & Judge, G. G. (1993) *Learning and Practicing Econometrics* (New York: Wiley).

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995) *Continuous Univariate Distributions*, vol. 2, 2nd edn (New York: Wiley).

Jørgesen, B. (1997) Proper dispersion models (with discussion), *Brazilian Journal of Probability and Statistics*, 11, pp. 89–140.

Koenker, R. (1981) A note on studentizing a test for heteroscedasticity, *Journal of Econometrics*, 17, pp. 107–112.

McCullagh, P. & Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn (London: Chapman and Hall).

Neter, J., Kutner, M. H., Nachtsheim, C. J. & Wasserman, W. (1996) *Applied Linear Statistical Models*, 4th edn (Chicago, IL: Irwin).

Nocedal, J. & Wright, S. J. (1999) *Numerical Optimization* (New York: Springer-Verlag).

Prater, N. H. (1956) Estimate gasoline yields from crudes, *Petroleum Refiner*, 35, pp. 236–238.

Rao, C. R. (1973) *Linear Statistical Inference and Its Applications*, 2nd edn (New York: Wiley).

Wei, B.-C., Hu, Y.-Q. & Fung, W.-K. (1998) Generalized leverage and its applications, *Scandinavian Journal of Statistics*, 25, pp. 25–37.

Appendix A

In this appendix we obtain the score function and the Fisher information matrix for (β, ϕ). The notation used here is defined in the second section. From equation (6) we get, for $i = 1, \ldots, k$,

$$\frac{\partial \ell(\beta, \phi)}{\partial \beta_i} = \sum_{t=1}^{n} \frac{\partial \ell_t(\mu_t, \phi)}{\partial \mu_t} \frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_t} \frac{\partial \eta_t}{\partial \beta_i} \tag{A1}$$

Note that $\mathrm{d}\mu_t/\mathrm{d}\eta_t = 1/g'(\mu_t)$. Also, from equation (7)

$$\frac{\partial \ell_t(\mu_t, \phi)}{\partial \mu_t} = \phi \left[\log \frac{y_t}{1-y_t} - \{\psi(\mu_t \phi) - \psi((1-\mu_t)\phi)\} \right] \tag{A2}$$

where $\psi(\cdot)$ is the digamma function, i.e. $\psi(z) = \mathrm{d} \log \Gamma(z)/\mathrm{d}z$ for $z > 0$. From regularity conditions, it is known that the expected value of the derivative in equation (7) equals zero, so that $\mu_t^* = \mathrm{E}(y_t^*)$, where y_t^* and μ_t^* are defined in the second section. Hence,

$$\frac{\partial \ell(\beta, \phi)}{\partial \beta_i} = \phi \sum_{t=1}^{n} (y_t^* - \mu_t^*) \frac{1}{g'(\mu_t)} x_{ti} \tag{A3}$$

We then arrive at the matrix expression for the score function for β given in equation (8). Similarly, it can be shown that the score function for ϕ can be written as in equation (9).

From equation (A1), the second derivative of $\ell(\beta, \phi)$ with respect to the βs is given by

$$\frac{\partial^2 \ell(\beta, \phi)}{\partial \beta_i \partial \beta_j} = \sum_{t=1}^{n} \frac{\partial}{\partial \mu_t} \left(\frac{\partial \ell_t(\mu_t, \phi)}{\partial \mu_t} \frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_t} \right) \frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_t} \frac{\partial \eta_t}{\partial \beta_j} x_{ti}$$

$$= \sum_{t=1}^{n} \left(\frac{\partial^2 \ell_t(\mu_t, \phi)}{\partial \mu_t^2} \frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_t} + \frac{\partial \ell_t(\mu_t, \phi)}{\partial \mu_t} \frac{\partial}{\partial \mu_t} \frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_t} \right) \frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_t} x_{ti} x_{tj}$$

Since $\mathrm{E}(\partial \ell_t(\mu_t, \phi)/\partial \mu_t) = 0$, we have

$$\mathrm{E}\left(\frac{\partial^2 \ell(\beta, \phi)}{\partial \beta_i \partial \beta_j} \right) = \sum_{t=1}^{n} \mathrm{E}\left(\frac{\partial^2 \ell_t(\mu_t, \phi)}{\partial \mu_t^2} \right) \left(\frac{\mathrm{d}\mu_t}{\mathrm{d}\eta_t} \right)^2 x_{ti} x_{tj}$$

Now, from equation (A2) we have

$$\frac{\partial^2 \ell_t(\mu_t, \phi)}{\partial \mu_t^2} = -\phi^2 \{ \psi'(\mu_t \phi) + \psi'((1 - \mu_t)\phi) \}$$

and hence

$$\mathrm{E}\left(\frac{\partial^2 \ell(\beta, \phi)}{\partial \beta_i \partial \beta_j} \right) = -\phi \sum_{t=1}^{n} w_t x_{ti} x_{tj}$$

In matrix form, we have that

$$\mathrm{E}\left(\frac{\partial^2 \ell(\beta, \phi)}{\partial \beta \partial \beta^{\mathsf{T}}} \right) = \phi X^{\mathsf{T}} W X$$

From equation (A3), the second derivative of $\ell(\beta, \phi)$ with respect to β_i and ϕ can be written as

$$\frac{\partial^2 \ell(\beta, \phi)}{\partial \beta_i \partial \phi} = \sum_{t=1}^{n} \left[(y_t^* - \mu_t^*) - \phi \frac{\partial \mu_t^*}{\partial \phi} \right] \frac{1}{g'(\mu_t)} x_{ti}$$

Since $\mathrm{E}(y_t^*) = \mu_t^*$ and $\partial \mu_t^* / \partial \phi = \psi'(\mu_t \phi)\mu_t - \psi'((1 - \mu_t)\phi)(1 - \mu_t)$, we arrive at

$$\mathrm{E}\left(\frac{\partial^2 \ell(\beta, \phi)}{\partial \beta_i \partial \phi} \right) = -\sum_{t=1}^{n} c_t \frac{1}{g'(\mu_t)} x_{ti}$$

In matrix notation, we then have

$$\mathrm{E}\left(\frac{\partial^2 \ell(\beta, \phi)}{\partial \beta \partial \phi}\right) = -X^{\top} T c$$

Finally, $\partial^2 \ell(\beta, \phi)/\partial \phi^2$ comes by differentiating the expression in equation (9) with respect to ϕ. We arrive at $\mathrm{E}(\partial^2 \ell(\beta, \phi)/\partial \phi^2) = -\Sigma_{t=1}^{n} d_t$, which, in matrix notation, can be written as

$$\mathrm{E}\left(\frac{\partial^2 \ell(\beta, \phi)}{\partial \phi^2}\right) = -\mathrm{tr}(D)$$

It is now easy to obtain the Fisher information matrix for (β, ϕ) given in equation (10).

Appendix B

In this Appendix, we show how to perform large sample inference in the beta regression model we propose. Consider, for instance, the test of the null hypothesis $\mathrm{H}_0: \beta_1 = \beta_1^{(0)}$ versus $\mathrm{H}_1: \beta_1 \neq \beta_1^{(0)}$, where $\beta_1 = (\beta_1, \ldots, \beta_m)^{\top}$ and $\beta_1^{(0)} = (\beta_1^{(0)}, \ldots, \beta_m^{(0)})^{\top}$, for $m < k$, and $\beta_1^{(0)}$ given. The log-likelihood ratio statistic is

$$\omega_1 = 2\{\ell(\hat{\beta}, \hat{\phi}) - \ell(\tilde{\beta}, \tilde{\phi})\}$$

where $\ell(\beta, \phi)$ is the log-likelihood function and $(\tilde{\beta}^{\top}, \tilde{\phi})^{\top}$ is the restricted maximum likelihood estimator of $(\beta^{\top}, \phi)^{\top}$ obtained by imposing the null hypothesis. Under the usual regularity conditions and under H_0, $\omega_1 \stackrel{\mathcal{D}}{\to} \chi_m^2$, so that a test can be performed using approximate critical values from the asymptotic χ_m^2 distribution.

In order to describe the score test, let $U_{1\beta}$ denote the m-vector containing the first m elements of the score function for β and let $K_{11}^{\beta\beta}$ be the $m \times m$ matrix formed out of the first m rows and the first m columns of K^{-1}. It can be shown, using equation (8), that $U_{1\beta} = \phi X_1^{\top} T(y^* - \mu^*)$, where X is partitioned as $[X_1 \; X_2]$ following the partition of β. Rao's score statistic can be written as

$$\omega_2 = \tilde{U}_{1\beta}^{\top} \tilde{K}_{11}^{\beta\beta} \tilde{U}_{1\beta}$$

where tildes indicate that the quantities are evaluated at the restricted maximum likelihood estimator. Under the usual regularity conditions and under H_0, $\omega_2 \stackrel{\mathcal{D}}{\to} \chi_m^2$.

Asymptotic inference can also be performed using Wald's test. The test statistic for the test of $\mathrm{H}_0: \beta_1 = \beta_1^{(0)}$ is

$$\omega_3 = (\hat{\beta}_1 - \beta_1^{(0)})^{\top} (\hat{K}_{11}^{\beta\beta})^{-1} (\hat{\beta}_1 - \beta_1^{(0)})$$

where $\hat{K}_{11}^{\beta\beta}$ equals $K_{11}^{\beta\beta}$ evaluated at the unrestricted maximum likelihood estimator, and $\hat{\beta}_1$ is the maximum likelihood estimator of β_1. Under mild regularity conditions and under H_0, $\omega_3 \stackrel{\mathcal{D}}{\to} \chi_m^2$. In particular, for testing the significance of the ith regression parameter (β_i), $i = 1, \ldots, k$, one can use the signed square root of Wald's statistic, i.e. $\hat{\beta}_i / \mathrm{se}(\hat{\beta}_i)$, where $\mathrm{se}(\hat{\beta}_i)$ is the asymptotic

standard error of the maximum likelihood estimator of $\hat{\beta}_i$ obtained from the inverse of Fisher's information matrix evaluated at the maximum likelihood estimates. The limiting null distribution of the test statistic is standard normal.

An approximate $(1-\alpha)\times 100\%$ confidence interval for β_i, $i=1,\ldots,k$ and $0<\alpha<1/2$, has limits given by $\hat{\beta}_i \pm \Phi^{-1}(1-\alpha/2)\mathrm{se}(\hat{\beta}_i)$. Additionally, approximate confidence regions for sets of regression parameters can be obtained by inverting one of the three large sample tests described above. Similarly, an asymptotic $(1-\alpha)\times 100\%$ confidence interval for ϕ has limits $\hat{\phi} \pm \Phi^{-1}(1-\alpha/2)\mathrm{se}(\hat{\phi})$, where $\mathrm{se}(\hat{\phi}) = \hat{\gamma}^{-1/2}$. Additionally, an approximate $(1-\alpha)\times 100\%$ confidence interval for the odds ratio $e^{c\beta_i}$, when the logit link is used, is

$$[\exp\{c(\hat{\beta}_i - \Phi^{-1}(1-\alpha/2)\mathrm{se}(\hat{\beta}_i))\}, \exp\{c(\hat{\beta}_i + \Phi^{-1}(1-\alpha/2)\mathrm{se}(\hat{\beta}_i))\}]$$

Finally, an approximate $(1-\alpha)\times 100\%$ confidence interval for μ, the mean of the response, for a given vector of covariate values x can be computed as

$$[g^{-1}(\hat{\eta} - \Phi^{-1}(1-\alpha/2)\mathrm{se}(\hat{\eta})),\ g^{-1}(\hat{\eta} + \Phi^{-1}(1-\alpha/2)\mathrm{se}(\hat{\eta}))]$$

where $\hat{\eta} = x^{\top}\hat{\beta}$ and $\mathrm{se}(\hat{\eta}) = \sqrt{x^{\top}\hat{\mathrm{cov}}(\hat{\beta})x}$; here, $\hat{\mathrm{cov}}(\hat{\beta})$ is obtained from the inverse of Fisher's information matrix evaluated at the maximum likelihood estimates by excluding the row and column of this matrix corresponding to the precision parameter. The above interval is valid for strictly increasing link functions.

Appendix C

Here we shall obtain approximations for w_t and μ_t^*, $t=1,\ldots,n$, when $\mu_t\phi$ and $(1-\mu_t)\phi$ are large. At the outset, note that (Abramowitz & Stegun, 1965, p. 259), as $z\to\infty$,

$$\psi(z) = \log(z) - \frac{1}{2z} - \frac{1}{12z^2} + \frac{1}{120z^4} + \ldots \tag{C1}$$

$$\psi'(z) = \frac{1}{z} + \frac{1}{2z^2} + \frac{1}{6z^3} - \frac{1}{30z^5} + \ldots \tag{C2}$$

In what follows, we shall drop the subscript t (that indexes observations). When $\mu\phi$ and $(1-\mu)\phi$ are large, it follows from equation (C2) that

$$w \approx \phi\left\{\frac{1}{\mu\phi} + \frac{1}{(1-\mu)\phi}\right\}\frac{1}{g'(\mu)^2} = \frac{1}{\mu(1-\mu)}\frac{1}{g'(\mu)^2}$$

Also, from equation (C1) we obtain

$$\mu^* \approx \log(\mu\phi) - \log((1-\mu)\phi) = \log\left(\frac{\mu}{1-\mu}\right)$$