

Bootstrap prediction intervals in beta regressions

Patrícia L. Espinheira · Silvia L. P. Ferrari ·
Francisco Cribari-Neto

Received: 21 September 2013 / Accepted: 18 March 2014 / Published online: 3 April 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract We address the issue of constructing prediction intervals for responses that assume values in the standard unit interval, $(0, 1)$. The response is modeled using the class of beta regression models and we introduce percentile and BC_a (bias-corrected and accelerated) bootstrap prediction intervals. We present Monte Carlo evidence on the finite sample behavior of such intervals. An empirical application is presented and discussed.

Keywords Beta distribution · Beta regression · Bootstrap · Prediction · Prediction interval

1 Introduction

The beta distribution is commonly used to model random variables that assume values in $(0, 1)$, such as rates, percentages and proportions. The beta density can display quite different shapes depending on the parameter values. Oftentimes the variable of interest is related to a set of independent (explanatory) variables. Ferrari and Cribari-Neto (2004) introduced a regression model in which the response is beta-distributed, its

P. L. Espinheira · F. Cribari-Neto
Departamento de Estatística, Universidade Federal de Pernambuco,
Cidade Universitária, Recife, PE 50740–540, Brazil
e-mail: patespipa@de.ufpe.br

F. Cribari-Neto
e-mail: cribari@de.ufpe.br

S. L. P. Ferrari (✉)
Departamento de Estatística, Universidade de São Paulo,
Rua do Matão, 1010, São Paulo, SP 05508–090, Brazil
e-mail: silviaferrari@usp.br

mean being related to a linear predictor through a link function. The linear predictor includes independent variables and regression parameters. Their model also includes a precision parameter whose reciprocal can be viewed as a dispersion measure. In the standard formulation of the beta regression model it is assumed that the precision is constant across observations. However, in many practical situations this assumption does not hold. [Smithson and Verkuilen \(2005\)](#), among others, consider a beta regression specification in which dispersion is not constant, but is a function of covariates and unknown parameters. Such models are known as varying dispersion beta regression models. Parameter estimation is carried out by maximum likelihood (ML) and standard asymptotic hypothesis testing can be easily performed. Practitioners can use the `betareg` package, which is available for the R statistical software (<http://www.r-project.org>), for fitting beta regressions. [Cribari-Neto and Zeileis \(2010\)](#) provide an overview of varying dispersion beta regression modeling using the `betareg` package.

An empirical application we shall address in this paper relates to the distribution of natural gas for home usage (e.g., in stoves, ovens and water heaters) in São Paulo, Brazil. Such a distribution is based on a simultaneity factor that assumes values in the standard unit interval, $(0, 1)$. It relates to the nominal power and to the number of appliances that use natural gas. Given these factors, the company that supplies the gas tries to forecast the probability of simultaneous appliances usage in order to decide how much gas to supply to a given residential unit. According to [Zerbinatti \(2008\)](#), in 2005 the Instituto de Pesquisas Tecnológicas (IPT) e a Companhia de Gás de São Paulo (COMGÁS) computed the simultaneity factor for a number of residences. [Zerbinatti \(2008\)](#) modeled such data using different regression models and concluded that the best performing model was the logit model that used the natural logarithm of the computed power indicator as a covariate. It is noteworthy that one has to be careful not to underestimate the simultaneity factor when making projections using an estimated regression model since that could cause a shortage of natural gas supply. The author shows that the beta regression model can underpredict the response; see [Zerbinatti \(2008, Figure 4.11.b\)](#). It is thus important to have at disposal *prediction intervals* that can be used with beta regressions. This is the motivation for our paper.

Our main goal is to propose and numerically evaluate bootstrap prediction intervals that can be used with the beta regression model. That is, we construct intervals for unobserved response values corresponding to a given set of covariate values. At the outset, we consider the percentile method as described by [Davison and Hinkley \(1997\)](#) for generalized linear models. We also consider a more refined prediction interval, namely: the BC_a (bias-corrected accelerated). We obtain the BC_a prediction interval for new response values. (Notice that we construct prediction intervals, not confidence intervals.) The finite sample performances of the bootstrap intervals are evaluated using Monte Carlo simulations. Finally, the empirical application briefly described above is addressed.

2 The bootstrap method

Let $x = (x_1, \dots, x_n)$ be a random sample from the random variable X whose distribution function is \mathbb{F} . Let $\theta = \iota(\mathbb{F})$ be the parameter that indexes the population and let

$\widehat{\theta} = S(x)$ be an estimator of θ . In the bootstrap method, one obtains, from the original sample x , a large number of pseudo-samples (bootstrap samples) $x^* = (x_1^*, \dots, x_n^*)$, computes the quantity of interest for each pseudo sample (i.e., $\widehat{\theta}^* = S(x^*)$), and then one uses the empirical distribution of $\widehat{\theta}^* = S(x^*)$ as an estimate of the distribution of $\widehat{\theta}$. Bootstrap sampling can be performed from the empirical distribution function of x , given by $\widehat{F}(t) = \#\{x_i \leq t\}/n, t \in \mathbb{R}$, or from F after replacing θ by $\widehat{\theta}$, a consistent parameter estimator. The former is known as the nonparametric bootstrap whereas the latter is known as the parametric bootstrap.

3 Beta regression prediction intervals

Let y_1, \dots, y_n be independent random variables such that each y_t , for $t = 1, \dots, n$, is beta distributed, i.e., each y_t has density function given by

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

where $0 < \mu < 1$ and $\phi > 0$. Here, $E(y) = \mu$ and $\text{var}(y) = V(\mu)/(1 + \phi)$, where $V(\mu) = \mu(1 - \mu)$. In the beta regression model introduced by Ferrari and Cribari-Neto (2004) the mean of y_t can be written as

$$g(\mu_t) = x_t^\top \beta = \eta_t. \quad (2)$$

In addition to the relation given in (2), it is possible to assume that the precision parameter is not constant and write

$$h(\phi_t) = z_t^\top \gamma = \vartheta_t. \quad (3)$$

In (2) and (3), η_t and ϑ_t are linear predictors, $\beta = (\beta_1, \dots, \beta_k)^\top$ and $\gamma = (\gamma_1, \dots, \gamma_q)^\top$ are unknown parameter vectors ($\beta \in \mathbb{R}^k; \gamma \in \mathbb{R}^q$), x_{t1}, \dots, x_{tk} and z_{t1}, \dots, z_{tq} are fixed covariates ($k + q < n$) and $g(\cdot)$ and $h(\cdot)$ are link functions, which are strictly increasing and twice-differentiable.

Fitted regression models are oftentimes used to predict out-of-sample response values. In the regression model described by (1)–(3), $x_t^\top = (x_{t1}, \dots, x_{tk})$ is a set of observed covariate values and y_t is the t th observed response. Let $x_+^\top = (x_{+1}, \dots, x_{+k})$ denote a new set of covariate values and let y_+ be the corresponding unobserved response value. The latter can be predicted by $\widehat{\mu}_+ = g^{-1}(\sum_{i=1}^k x_{+i} \widehat{\beta}_i)$, where $\widehat{\beta}_i$ is the maximum likelihood estimate of β_i computed using the original sample. It is useful to obtain a *prediction interval*, which is given by lower and upper limits that are statistics associated with a given desired coverage level. In what follows, we shall construct such an interval using estimates of the prediction error distribution.

3.1 Percentile prediction intervals for beta regressions

Let $\mathcal{R}(y, \mu)$ denote a monotonic function of y that has constant variance for all observations. Assume that the mean μ_+ and the distribution of $\mathcal{R}(y, \mu)$ are known, and

denote the α th quantile of such a distribution by δ_α ($0 < \alpha < 1/2$). The limits of the $1 - \alpha$ prediction interval has lower and upper limits given by $y_{+, \alpha/2}$ and $y_{+, 1-\alpha/2}$, respectively, which satisfy $\mathcal{R}(y_{+, \alpha/2}, \mu_+) = \delta_{\alpha/2}$ and $\mathcal{R}(y_{+, 1-\alpha/2}, \mu_+) = \delta_{(1-\alpha/2)}$, where $\mathcal{R}(y, \mu_+)$ is the prediction error. If $\widehat{\mu}$, the estimate of μ , is computed independently from y_+ and the quantiles of $\mathcal{R}(y_+, \widehat{\mu})$ are known, the prediction interval can be easily computed. The distribution of $\mathcal{R}(y_+, \widehat{\mu})$, however, is typically unknown. In what follows, we shall use data resampling to estimate it and then obtain the desired quantiles from such an estimated distribution, which are used to construct the prediction interval. In the data resampling mechanism we shall work with a normalized version of $\mathcal{R}(y, \widehat{\mu})$ whose distribution has constant variance.

For the beta regression model, we consider

$$\mathcal{R}(y, \mu) = \frac{y_t^* - \mu_t^*}{\sqrt{v_t}}, \quad (4)$$

where $E(y_t^*) = \mu_t^*$ and $\text{var}(y_t^*) = v_t$, with $v_t = \psi'(\mu_t \phi_t) + \psi'((1 - \mu_t)\phi_t)$. Here,

$$y_t^* = \log\{y_t/(1 - y_t)\} \quad \text{and} \quad \mu_t^* = \psi(\mu_t \phi_t) - \psi((1 - \mu_t)\phi_t). \quad (5)$$

Hence, $\mathcal{R}(y, \mu)$ is a monotonic function of y with zero mean and unit variance [see Ferrari et al. 2008, Eq. (8)]. In the data resampling mechanism we shall use the standardized version of $\mathcal{R}(y, \widehat{\mu})$ given by

$$r_t = \frac{y_t^* - \widehat{\mu}_t^*}{\sqrt{\widehat{v}_t(1 - h_{tt}^*)}}, \quad (6)$$

which was proposed in Espinheira et al. (2011) and is known as standardized weighted residual 2. It is a standardized residual obtained using Fisher's scoring iterative algorithm for β under varying dispersion. Here, h_{tt}^* is the t th diagonal element of

$$H^* = (\widehat{W}\widehat{V})^{1/2} X(X^\top \widehat{V}\widehat{W}X)^{-1} X^\top (\widehat{V}\widehat{W})^{1/2},$$

where X is the $n \times k$ matrix of covariates ($k < n$), $W = \text{diag}\{w_1, \dots, w_n\}$ with $w_t = \phi_t v_t / \{g'(\mu_t)\}^2$ and

$$V = \text{diag}\{\phi_1, \dots, \phi_n\}. \quad (7)$$

Using the approach outlined by Davison and Hinkley (1997, p. 340) for generalized linear models we can construct the $1 - \alpha$ percentile prediction interval using the $\alpha/2$ and $1 - \alpha/2$ quantiles of \widehat{G} , the bootstrap approximation to $\mathcal{R}(y_+, \widehat{\mu})$.

It is well known that percentile confidence interval for θ , the parameter that indexes a given population, can display poor behavior in small samples when based on a highly biased estimator of that parameter (DiCiccio and Tibshirani 1987). A more refined approach is known as BC_a (bias-corrected and accelerated). It accounts for bias and for the fact that the estimator standard error may vary with θ .

3.2 BC_α confidence intervals

Assume that there exists a function $h(\theta) = \rho$, which is monotonic decreasing, and constants a and v_0 such that $\widehat{\rho} - \rho \sim \mathcal{N}(-v_0(1+a\rho), (1+a\rho)^2)$. The exact upper limit of the $1 - \alpha$ confidence interval for ρ is $\rho[\alpha] = \widehat{\rho} + \text{se}_{\widehat{\rho}}\{v_0 + z_\alpha\}/\{1 - a(v_0 + z_\alpha)\}$, where z_α is the α standard normal quantile. Let P denote the distribution function of $\widehat{\theta}$. Using the inverse transformation $h^{-1}(\cdot)$, we obtain an estimate of α :

$$\tilde{\alpha} = \Phi\left(\widehat{v}_0 + \frac{\widehat{v}_0 + z_\alpha}{1 - \widehat{a}(\widehat{v}_0 + z_\alpha)}\right), \tag{8}$$

where Φ is the standard normal distribution function. The $1 - \alpha$ BC_α confidence interval is given by the $\tilde{\alpha}/2$ and $1 - \tilde{\alpha}/2$ quantiles of \widehat{G} when z_α in (8) is replaced by $z_{\alpha/2}$ and $z_{1-\alpha/2}$, respectively.

The constant v_0 accounts for any bias of the plug-in estimator. According to Efron (1987) it can be estimated as

$$\widehat{v}_0 = \Phi^{-1}\left(\frac{\#\widehat{\theta}_b < \widehat{\theta}}{B}\right), \tag{9}$$

where $\widehat{\theta}_b$ is the b th bootstrap estimate of θ and B is the number of bootstrap replications. Roughly speaking, \widehat{v}_0 measures the discrepancy between the median of $\widehat{\theta}^*$ and $\widehat{\theta}$, in normal units. If $\widehat{v}_0 = 0$, then $\widehat{\theta} = \text{median}(\widehat{\theta}^*)$ and the bias correction is not needed.

The acceleration constant a accounts for the rate of change in the standard error of $\widehat{\theta}$ with respect to θ . According to Efron (1987), in one parameter models a good approximation to a is

$$a \approx \frac{1}{6} \text{skew}_{\theta=\widehat{\theta}}(\dot{\ell}_\theta), \tag{10}$$

where $\dot{\ell}_\theta = \dot{\ell} = d \log f(y; \theta)/d\theta$. Note that the skewness of $\dot{\ell}$, $E[(\dot{\ell} - E[\dot{\ell}])^3]/\{E[(\dot{\ell} - E[\dot{\ell}])^2]\}^{3/2}$, is evaluated at $\widehat{\theta}$. According to Davison and Hinkley (1997) the expression in (10) is equivalent to

$$a \approx \frac{1}{6} \frac{E[\dot{\ell}(\widehat{\theta})^3]}{\text{var}[\dot{\ell}(\widehat{\theta})]^{3/2}}, \tag{11}$$

which can be estimated using data resampling.

3.3 BC_α prediction intervals for the realization of a random variable

Our goal in this paper, however, is not to construct confidence intervals for a parameter, but to construct *prediction intervals* for the realization of a random variable. In that sense, we shall now consider the proposal made in Mojirsheibani and Tibshirani (1996). The authors develop prediction intervals for $\widehat{\theta}_+$, an efficient estimator of a scalar parameter θ , based on a new sample of size n^+ .

Based on expressions (8)–(10), the authors suggest the following choice of value for v_0 :

$$\widehat{v}_0 = \Phi^{-1} \left(\frac{\#\widehat{\theta}_{+,b} < \widehat{\theta}_n}{B} \right), \tag{12}$$

where $\widehat{\theta}_n$ and $\widehat{\theta}_+$ are the MLEs of θ obtained using y_n and y_+ , respectively, $y_n = (y_1, \dots, y_n)^\top$ being the original sample and $y_+ = (y_{1+}, \dots, y_{n+})^\top$ being the new sample values. Additionally, they suggest using

$$a = \frac{1}{6} \text{skew}_{\theta=\widehat{\theta}_n}(\dot{\ell}_\theta(\widehat{\theta}_+)) \tag{13}$$

with a multiplicative correction of $(n/n^+)^{-1/2}$.

3.4 BC_a prediction intervals for new and unobserved response values

The BC_a scheme presented Sect. 3.3 can be used to construct prediction intervals for $\widehat{\theta}_+$ under new response values. Our goal, nonetheless, is to construct prediction intervals for new, unobserved response values given a set of covariates values. The BC_a method we propose is a modification of the existing BC_a methods (Sects. 3.2 and 3.3). In particular, we obtain expressions for a and v_0 when the interest lies in predicting a given observation. This is the main difference between our result and the existing results and is also our main contribution to the literature.

Based on (10), we propose using

$$a = \frac{1}{6} \text{skew}_{\theta=\widehat{\theta}_n}(\dot{\ell}(\widehat{\mu}_+)) \approx \frac{1}{6} \frac{E[\dot{\ell}_+^3]}{\text{var}[\dot{\ell}_+]^{3/2}}, \tag{14}$$

which is to be corrected by the multiplicative factor given by $(n/n^+)^{-1/2}$. It can be shown that, in the class of beta regressions, $\dot{\ell}_t = d \log f(y; \mu, \phi) / d\mu = \phi_t(y_t^* - \mu_t^*)$. Thus, $\dot{\ell}_+ = \phi_+(y_+^* - \mu_+^*)$, which becomes $\dot{\ell}_+ = \phi(y_+^* - \mu_+^*)$ under constant dispersion.

Estimation of a can be performed using data resampling, e.g., bootstrap. We used that approach in a numerical experiment, i.e., we used the bootstrap to obtain estimates of $E[\dot{\ell}_t^3]$ and $\text{var}[\dot{\ell}_t]$, which in turn allowed us to obtain an estimate for a . Our numerical results showed, however, that this approach did not display good small sample performance. A better performing approach proved to be the one in which a is calculated analytically. For the varying dispersion beta regression model we obtain, after some algebra, $E[\dot{\ell}_t^3] = \varphi_t = \phi_t^3 \{\psi''(\mu_t, \phi_t) - \psi''((1 - \mu_t)\phi_t)\}$ and $\text{var}[\dot{\ell}_t] = v_t = \phi_t^2 \{\psi'(\mu_t, \phi_t) + \psi'((1 - \mu_t)\phi_t)\}$. It then follows that $E[\dot{\ell}_+^3] = \varphi_+$, $\text{var}[\dot{\ell}_+] = v_+$ and $\widehat{a} = (1/6)\widehat{\varphi}_+/\widehat{v}_+^{3/2}$, with $\widehat{\varphi}_\pm = \widehat{\phi}_\pm^3 \{\psi''(\widehat{\mu}_\pm \widehat{\phi}_\pm) - \psi''((1 - \widehat{\mu}_\pm)\widehat{\phi}_\pm)\}$ and $\widehat{v}_\pm = \widehat{\phi}_\pm^2 \{\psi'(\widehat{\mu}_\pm \widehat{\phi}_\pm) + \psi'((1 - \widehat{\mu}_\pm)\widehat{\phi}_\pm)\}$. Here,

$$\widehat{\phi}_+ = h^{-1} \left(\sum_{j=1}^q z_{+,j} \widehat{\gamma}_j \right). \tag{15}$$

In (15), $\widehat{\gamma}_j$ is the MLE of γ_j and $z_{+,j}$ is the j th component of $z_+^\top = (z_{+1}, \dots, z_{+q})$, i.e., it is the j th component of the set of dispersion regressors associated with the unobserved response y_+ .

Next, we shall propose an estimator for v_0 . Since our interest lies in the estimation of the distribution of $\mathcal{R}(y_+, \widehat{\mu})$ we shall use an estimate which is based on such quantity. Our proposal is to use

$$\widehat{v}_0 = \Phi^{-1}\left(\frac{\#\mathcal{R}_{a_+,b} < \mathcal{R}_m}{B}\right), \tag{16}$$

where \mathcal{R}_m is the median of $\mathcal{R}_1, \dots, \mathcal{R}_n$, which is computed using the original data and (4), and $\mathcal{R}_{a_+,b}$ is defined in algorithm given below; see Eq. (17).

3.5 Algorithm

Our algorithm can be outlined as follows. It is intended for n_1 predictions and uses B bootstrap replications. The bootstrap replications are indexed as $b = 1, \dots, B$.

1. For $t = 1, \dots, n$, randomly draw $r_{t,b}$ from r_1, \dots, r_n (with replacement).
2. Construct a bootstrap sample (y_b, X, Z) , where $y_b = (y_{1,b}, \dots, y_{n,b})^\top$, such that

$$y_{t,b} = \frac{\exp(\widehat{\mu}_t^* + r_{t,b}\sqrt{\widehat{v}_t})}{1 + \exp(\widehat{\mu}_t^* + r_{t,b}\sqrt{\widehat{v}_t})}$$

is obtained as the solution to $\mathcal{R}(y_t, \widehat{\mu}_t) = r_{t,b}$.

3. Using (y_b, X, Z) compute $\widehat{\beta}_b$ and $\widehat{\gamma}_b$, the bootstrap estimates of β and γ , respectively. Here, Z is the $n \times q$ matrix of covariates used in the dispersion submodel. Using the matrices of new observations on the regressors, X_+ ($n_1 \times k$) and Z_+ ($n_1 \times q$), together with $\widehat{\beta}_b$ and $\widehat{\gamma}_b$, obtain $\widehat{\mu}_{+,b}, \widehat{\phi}_{+,b}, \widehat{\mu}_{+,b}^*, \widehat{v}_{+,b}$, which are n_1 -vectors.
4. For each new observation $a_+ = 1, \dots, n_1$:
 - (a) Randomly draw $r_{a_+,b}$ from r_1, \dots, r_n .
 - (b) Compute

$$y_{a_+,b} = \frac{\exp(\widehat{\mu}_{a_+,b}^* + r_{a_+,b}\sqrt{\widehat{v}_{a_+,b}})}{1 + \exp(\widehat{\mu}_{a_+,b}^* + r_{a_+,b}\sqrt{\widehat{v}_{a_+,b}})}$$

- (c) Compute the prediction error

$$\mathcal{R}_{a_+,b}(y_{a_+,b}, \widehat{\mu}_{a_+,b}^*) = \frac{y_{a_+,b}^* - \widehat{\mu}_{a_+,b}^*}{\sqrt{\widehat{v}_{a_+,b}}}, \tag{17}$$

where $y_{a_+,b}^* = \log \{y_{a_+,b}/(1 - y_{a_+,b})\}$. For each new observation, sort the B values $\mathcal{R}_{a_+,b}$, such that $\mathcal{R}_{a_+,b(1)} \leq \dots \leq \mathcal{R}_{a_+,b(B)}$. Compute the percentile quantiles

$$\delta_{P_{a^+}^{a^+}(\alpha/2)}^* = \mathcal{R}_{a_+(B(\alpha/2))} \text{ and } \delta_{P_{a^+}^{a^+}(1-\alpha/2)}^* = \mathcal{R}_{a_+(B(1-\alpha/2))},$$

and the BC_a quantiles

$$\delta_{BC_{a^+}^{a^+}(\alpha/2)}^* = \mathcal{R}_{a_+(B(\tilde{\alpha}/2))} \text{ and } \delta_{BC_{a^+}^{a^+}(1-\alpha/2)}^* = \mathcal{R}_{a_+(B(1-\tilde{\alpha}/2))},$$

with $\tilde{\alpha}/2$ given in (8). Finally, obtain the prediction interval limits, percentile ($\delta_{a^+}^* = \delta_{P_{a^+}^{a^+}}^*$) or BC_a ($\delta_{a^+}^* = \delta_{BC_{a^+}^{a^+}}^*$), using

$$y_{a^+,I} = \frac{\exp(\widehat{\mu}_{a^+}^* + \delta_{a^+(\alpha/2)}^* \sqrt{\widehat{v}_{a^+}})}{1 + \exp(\widehat{\mu}_{a^+}^* + \delta_{a^+(\alpha/2)}^* \sqrt{\widehat{v}_{a^+}})}$$

$$y_{a^+,S} = \frac{\exp(\widehat{\mu}_{a^+}^* + \delta_{a^+(1-\alpha/2)}^* \sqrt{\widehat{v}_{a^+}})}{1 + \exp(\widehat{\mu}_{a^+}^* + \delta_{a^+(1-\alpha/2)}^* \sqrt{\widehat{v}_{a^+}})}.$$

Here, $\widehat{\mu}_{a^+}^*$ and \widehat{v}_{a^+} are the quantities μ^* and v evaluated at $\widehat{\mu}_{a^+} = g^{-1}(x_{a^+}^\top \widehat{\beta})$ and $\widehat{\phi}_{a^+} = h^{-1}(z_{a^+}^\top \widehat{\gamma})$, $x_{a^+}^\top$ and $z_{a^+}^\top$ being the a_+ -th rows of X_+ and Z_+ relative to the new observations, respectively, $a_+ = 1, \dots, n_1$. The values $y_{a^+,I}$ and $y_{a^+,S}$ are obtained, respectively, as the solutions to $\mathcal{R}(y_{a^+}, \widehat{\mu}_{a^+}) = \delta_{a^+(\alpha/2)}^*$ and $\mathcal{R}(y_{a^+}, \widehat{\mu}_{a^+}) = \delta_{a^+(1-\alpha/2)}^*$.

4 Simulation results

The simulation results presented in this section were obtained using both fixed and varying dispersion beta regressions as data generating processes. Table 1 contains numerical results for the fixed dispersion beta regression model given by

$$\log\left(\frac{\mu_t}{1 - \mu_t}\right) = \beta_1 + \beta_2 x_{t2}.$$

The sample sizes are $n = 40, 80, 120$ and the precisions are $\phi = 50, 150, 400$. There are five different scenarios. In the first three scenarios, $\mu \in (0.15, 0.80)$, $\mu \in (0.95, 0.98)$ and $\mu \in (0.02, 0.07)$; the covariate values were generated from the standard normal distribution. In the remaining two scenarios we generated the covariate values from the t_3 and unit mean exponential distributions in order to introduce leverage points in the data. The number of Monte Carlo replications is 5,000 and for each replication we perform $B = 500$ bootstrap replications. The nominal coverage of all intervals is 95%. The figures in Table 1 are empirical coverages (%).

The numerical results presented in Table 1 show that the BC_a and percentile intervals perform similarly, the BC_a outperforming the percentile method in some situations. For instance, when $\mu \in (0.02, 0.07)$, $\phi = 50$ and $n = 40$, the percentile coverage equals

Table 1 Empirical coverages (%). The model is $\log(\mu_t/(1 - \mu_t)) = \beta_1 + \beta_2 x_{t2}$, $B = 500$ and the nominal coverage is 95 %

μ	$\mu \in (0.15, 0.80)$								
ϕ	50			150			400		
n	40	80	120	40	80	120	40	80	120
Percentile	97.9	96.8	97.2	92.4	95.4	94.7	94.8	96.3	94.7
BC_a	98.0	96.1	97.3	93.8	95.6	97.1	94.8	96.3	94.7
μ	$\mu \in (0.95, 0.98)$								
ϕ	50			150			400		
n	40	80	120	40	80	120	40	80	120
Percentile	96.0	90.9	96.5	92.0	97.0	95.5	97.9	94.7	97.3
BC_a	95.8	90.5	95.9	91.2	94.4	95.5	95.5	94.7	96.9
μ	$\mu \in (0.02, 0.07)$								
ϕ	50			150			400		
n	40	80	120	40	80	120	40	80	120
Percentile	87.3	96.5	96.9	90.3	90.5	90.3	89.6	92.6	94.7
BC_a	97.5	98.8	99.6	90.8	90.5	90.3	89.8	93.9	94.7
μ	Covariate values generated from t_3								
ϕ	50			150			400		
n	40	80	120	40	80	120	40	80	120
Percentile	98.8	96.3	93.2	94.6	93.2	94.5	93.5	96.5	96.6
BC_a	98.1	95.9	92.3	94.6	90.9	97.0	92.8	96.1	96.7
μ	Covariate values generated from $\exp(1)$								
ϕ	50			150			400		
n	40	80	120	40	80	120	40	80	120
Percentile	97.2	90.1	94.8	95.8	91.7	95.0	94.7	96.2	92.2
BC_{a3}	99.5	96.3	95.9	95.8	95.1	95.1	97.5	96.2	92.2

87.3 % whereas the BC_a coverage is 97.5 %. We also note that when the covariate values are generated from the exponential distribution, the BC_a is consistently superior to the percentile method. For example, when $\phi = 50$ and $n = 80$, the BC_a and percentile coverages are 96.3 and 90.1 %, respectively. There are some situations, however, in which the percentile outperforms the BC_a , such as when $\phi = 150$ and $n = 80$; their respective coverages are 93 and 90 %. It is also noteworthy that the finite sample performances of both interval estimators improve when the sample size increases and also when the value of the precision parameter increases.

We have also carried out Monte Carlo simulations using a varying dispersion beta regression model. The data generating process is

$$\log\left(\frac{\mu_t}{1 - \mu_t}\right) = \beta_1 + \beta_2 x_{t2} \text{ and } \log(\phi_t) = \gamma_1 + \gamma_2 x_{t2}.$$

We measure the intensity of nonconstant dispersion as $\lambda = \phi_{\max}/\phi_{\min}$ and report results for $\lambda = 20, 50, 100$. The empirical coverages are given in Table 2. The results

Table 2 Empirical coverages (%)

μ	$\mu \in (0.15, 0.80)$								
λ	20			50			100		
n	40	80	120	40	80	120	40	80	120
Percentile	90.8	97.1	96.4	95.3	93.9	94.8	92.3	96.1	94.9
BC_a	90.8	97.1	97.1	92.6	93.9	94.8	92.3	96.1	95.5
μ	$\mu \in (0.95, 0.98)$								
λ	20			50			100		
n	40	80	120	40	80	120	40	80	120
Percentile	98.3	93.8	92.3	97.7	95.4	95.2	92.1	96.1	95.3
BC_a	94.8	93.8	90.7	97.0	95.4	95.1	92.1	96.1	94.8
μ	$\mu \in (0.02, 0.08)$								
λ	20			50			100		
n	40	80	120	40	80	120	40	80	120
Percentile	90.7	94.1	94.3	89.9	90.7	91.5	92.1	96.1	95.3
BC_a	90.7	94.2	95.3	91.3	94.8	93.4	92.1	96.1	94.8
μ	Covariate values generated from t_3								
λ	20			50			100		
n	40	80	120	40	80	120	40	80	120
Percentile	94.2	95.4	96.6	93.6	95.5	94.9	93.4	95.9	93.9
BC_a	91.6	95.4	96.6	91.7	95.6	94.9	93.4	94.4	92.1

The model is $\log(\mu_t/(1 - \mu_t)) = \beta_1 + \beta_2x_{t2}$ and $\log(\phi_t) = \gamma_1 + \gamma_2x_{t2}$, $B = 500$ and the nominal coverage is 95 %

show that the empirical coverages are sensitive to the intensity of nonconstant dispersion. Overall, the two methods are competitive. The percentile method outperforms the BC_a when the data include leverage points (covariate values obtained as random draws from the t_3 distribution). For instance, when $\lambda = 20$ and $n = 40$, their coverages are 94.2 and 91.6 %, respectively. On the other hand, the BC_a outperforms the percentile method when $\mu \in (0.02, 0.07)$. For example, when $\lambda = 50$ and $n = 80$, the respective coverages are 94.8 and 90.7 %.

We performed additional simulations in which we increased the number of covariates, used different covariates in the mean and precision submodels and incorrectly estimated a fixed dispersion beta regression when the true data generating process had varying dispersion. The new numerical results are presented in Tables 3, 4 and 5. We considered the following beta regression model:

$$\log\left(\frac{\mu_t}{1 - \mu_t}\right) = \beta_1 + \beta_2x_{t2} + \beta_3x_{t3} + \beta_4x_{t4} + \beta_5x_{t5}, \tag{18}$$

$$\log(\phi_t) = \gamma_1 + \gamma_2z_{t2} + \gamma_3z_{t3} + \gamma_4z_{t4} + \gamma_5z_{t5}. \tag{19}$$

The covariate values in the mean submodel were obtained as random draws from the standard uniform distribution whereas those in the precision submodel were obtained

Table 3 Empirical coverages (%)

β	$\beta_1 = 1.7, \beta_2 = -2, 5\beta_3 = 2.0, \beta_4 = 0.0, \beta_5 = 0.0 \quad \mu \in (0.20, 0.90)$								
ϕ	50			150			400		
n	40	80	120	40	80	120	40	80	120
Percentile	91.6	97.5	97.3	93.5	92.9	92.5	94.0	94.9	94.2
BC_a	98.6	97.6	97.3	93.5	92.9	92.8	94.0	94.6	94.72
β	$\beta_1 = 2.0, \beta_2 = -1.0, \beta_3 = 3.0, \beta_4 = -2.0, \beta_5 = 0.0 \quad \mu \in (0.18, 0.90)$								
ϕ	50			150			400		
n	40	80	120	40	80	120	40	80	120
Percentile	93.6	97.9	92.1	91.6	97.6	96.0	92.7	94.3	94.8
BC_a	97.5	99.3	97.9	91.6	97.6	96.0	92.9	94.3	94.8
β	$\beta_1 = 2.0, \beta_2 = -1.0, \beta_3 = 3.0, \beta_4 = -2.0, \beta_5 = 1.0 \quad \mu \in (0.17, 0.91)$								
ϕ	50			150			400		
n	40	80	120	40	80	120	40	80	120
Percentile	91.5	95.7	91.2	92.7	96.2	95.7	93.7	94.5	95.1
BC_a	97.6	95.7	95.0	92.7	96.2	95.7	93.70	95.5	95.1

The model is $\log(\mu_t / (1 - \mu_t)) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 x_{t5}$, $B = 500$ and the nominal coverage is 95 %

as random draws from the $\mathcal{U}(-0.5, 0.5)$ distribution. Thus, the covariate values in the two submodels are not the same. The results in Table 3 were obtained using $\log(\phi_t) = \gamma_1$ (constant dispersion). The results in Table 4 were obtained using different values for the β 's and γ 's which lead to models with different number of covariates. Additionally, $\lambda = 20, 50, 100$. Finally, we considered the case in which the true data generating process has varying dispersion but a fixed dispersion beta regression is estimated; see Table 5. The results in Table 3 show that the intervals do not become considerably less accurate when the number of covariates increases, especially when the precision parameter equals 150 or 400. When $\phi = 50$ and $n = 40$ the percentile interval displays smaller coverage. The BC_a finite sample performance is not altered. The average interval lengths are also not substantially affected. The simultaneous increase in the number of covariates in the two submodels also does not noticeably affect the intervals finite sample performances; see Table 4.

The intervals finite sample behavior slightly change when the true data generating process has varying dispersion, but a fixed dispersion beta regression model is estimated. The results are presented in Table 5. The most extreme change takes place when $\lambda = 50$ and $n = 40$. For the correctly specified one-covariate model the percentile and BC_a coverage rates (average lengths) are 95.3 and 92.6 % (0.16 e 0.15), respectively. When the fixed dispersion model is estimated, these coverage rates (average lengths) become 100.0 % and 100.0 % (0.36 e 0.20). Overall, the BC_a average lengths are smaller than the percentile average lengths. Here, the coverage rates tend to decrease when the number of covariates increases. For instance, when there are three covariates, $\lambda = 20$ and $n = 80$ the percentile coverage (average length) becomes 88.4% (0.33).

Table 4 Empirical coverages (%)

β	$\beta_1 = 2.0, \beta_2 = -1.0, \beta_3 = 3.0, \beta_4 = 0.0, \beta_5 = 0.0 \quad \mu \in (0.50, 0.96)$								
γ	$\gamma_1 = 5.5, \gamma_2 = 2.5,$			$\gamma_1 = 6.0, \gamma_2 = 3.7,$			$\gamma_1 = 4.0, \gamma_3 = 4.5,$		
	$\gamma_3 = -2.0, \gamma_4 = 0.0,$			$\gamma_3 = -2.0, \gamma_4 = 0.0,$			$\gamma_3 = 1.0, \gamma_4 = 0.0,$		
	$\gamma_5 = 0.0$			$\gamma_5 = 0.0$			$\gamma_5 = 0.0$		
λ	20			50			100		
n	40	80	120	40	80	120	40	80	120
Percentile	96.0	98.7	99.1	95.4	93.0	95.5	92.4	98.2	98.9
BC _a	95.5	98.7	99.1	95.4	93.0	93.7	92.4	98.2	98.9
β	$\beta_1 = 2.0, \beta_2 = -1.0, \beta_3 = 3.0, \beta_4 = -2.0, \beta_5 = 0.0 \quad \mu \in (0.19, 0.90)$								
γ	$\gamma_1 = 2.0, \gamma_2 = 2.0,$			$\gamma_1 = 3.0, \gamma_2 = 2.0,$			$\gamma_1 = 4.0, \gamma_2 = 2.7,$		
	$\gamma_3 = 2.0, \gamma_4 = 1.0,$			$\gamma_3 = 3.0, \gamma_4 = 1.0,$			$\gamma_3 = 2.7, \gamma_4 = 2.0,$		
	$\gamma_5 = 0.0$			$\gamma_5 = 0.0$			$\gamma_5 = 0.0$		
λ	20			50			100		
n	40	80	120	40	80	120	40	80	120
Percentile	97.1	96.2	97.9	99.7	92.3	96.6	92.4	95.3	96.8
BC _a	97.1	96.2	98.3	99.7	92.3	96.6	92.4	94.0	98.2
β	$\beta_1 = 2.0, \beta_2 = -1.0, \beta_3 = 3.0, \beta_4 = -2.0, \beta_5 = 1.5 \quad \mu \in (0.16, 0.92)$								
γ	$\gamma_1 = 3.0, \gamma_2 = 1.5$			$\gamma_1 = 4.0, \gamma_2 = 2.5,$			$\gamma_1 = 4.0, \gamma_2 = 3.5,$		
	$\gamma_3 = -1.0, \gamma_4 = 1.0,$			$\gamma_3 = 3.0, \gamma_4 = 1.0,$			$\gamma_3 = -1.0, \gamma_4 = 1.0,$		
	$\gamma_5 = 2.0$			$\gamma_5 = 2.2$			$\gamma_5 = 2.2$		
λ	20			50			100		
n	40	80	120	40	80	120	40	80	120
Percentile	93.1	92.9	91.2	92.0	93.3	96.5	97.3	97.4	97.6
BC _a	93.1	92.9	91.2	92.0	93.5	96.5	97.3	97.5	97.5

The model is $\log(\mu_t / (1 - \mu_t)) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 x_{t5}$ and $\log(\phi_t) = \gamma_1 + \gamma_2 z_{t2} + \gamma_3 z_{t3} + \gamma_4 z_{t4} + \gamma_5 z_{t5}$, $B = 500$ and the nominal coverage is 95%

5 Empirical application

We shall now return to the application briefly described in the Introduction. Recall that it relates to the distribution of natural gas for home usage in São Paulo, Brazil. The distribution of natural gas is based on a simultaneity factor that assumes values in the standard unit interval, (0, 1).

Using the simultaneity factor one obtains the release indicator, i.e., an indicator of gas release in a given tubulation section: $Q_p = F \times Q_{max}$, where Q_p is the release, F is the simultaneity factor and Q_{max} is the maximum possible release. F assumes values in (0, 1), and can be interpreted as the ratio between effective and maximal intensities. We note that overpredictions of the simultaneity factor leads to excess supply of gas and, as consequence, inefficient allocation and higher costs.

According to [Zerbinatti \(2008\)](#), the Instituto de Pesquisas Tecnológicas (IPT) and the Companhia de Gás de São Paulo (COMGÁS) performed an extensive study in which data were collected in order to build a database on simultaneity factors and

Table 5 Empirical coverages (%)

β	$\beta_1 = 2.0, \beta_2 = -1.0,$ $\beta_3 = 0.0, \beta_4 = 0.0$ $\mu \in (0.66, 0.89)$			$\beta_1 = 2.0, \beta_2 = -1.0,$ $\beta_3 = 3.0, \beta_4 = 0.0$ $\mu \in (0.50, 0.96)$			$\beta_1 = 2.0, \beta_2 = -1.0,$ $\beta_3 = 3.0, \beta_4 = -2.0$ $\mu \in (0.19, 0.90)$		
λ	20								
γ	$\gamma_1 = 3.0, \gamma_2 = 3.0,$ $\gamma_3 = 0.0, \gamma_4 = 0.0$			$\gamma_1 = 3.0, \gamma_2 = 3.0,$ $\gamma_3 = 1.0, \gamma_4 = 0.0$			$\gamma_1 = 3.0, \gamma_2 = 2.0,$ $\gamma_3 = -1.5, \gamma_4 = 1.0$		
n	40	80	120	40	80	120	40	80	120
Percentile	100.0	100.0	100.0	100.0	100.0	100.0	97.4	88.4	96.0
BC_a	99.1	98.2	99.5	99.9	99.9	100.0	97.4	96.3	97.0
λ	50								
γ	$\gamma_1 = 3.0, \gamma_2 = 4.0,$ $\gamma_3 = 0.0, \gamma_4 = 0.0$			$\gamma_1 = 3.0, \gamma_2 = 3.0,$ $\gamma_3 = 2.0, \gamma_4 = 0.0$			$\gamma_1 = 3.0, \gamma_2 = 3.0,$ $\gamma_3 = 2.0, \gamma_4 = -1.0$		
n	40	80	120	40	80	120	40	80	120
Percentile	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
BC_a	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
λ	100								
γ	$\gamma_1 = 3.0, \gamma_2 = 5.0,$ $\gamma_3 = 0.0, \gamma_4 = 0.0$			$\gamma_1 = 3.0, \gamma_2 = 4.8,$ $\gamma_3 = -1.0, \gamma_4 = 0.0$			$\gamma_1 = 3.0, \gamma_2 = 3.0,$ $\gamma_3 = 2.0, \gamma_4 = -2.0$		
n	40	80	120	40	80	120	40	80	120
Percentile	100.0	100.0	100.0	100	100.0	100.0	92.3	96.1	94.9
BC_a	100.0	100.0	100.0	100	99.8	100.0	92.3	96.1	95.5

The true model is $\log(\mu_t/(1 - \mu_t)) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4}$ and $\log(\phi_t) = \gamma_1 + \gamma_2 z_{t2} + \gamma_3 z_{t3} + \gamma_4 z_{t4}$, but the postulated model is $\log(\phi_t) = \gamma_1$, $B = 500$ and the nominal coverage is 95 %

the corresponding maximal releases (computed power). The sampled households were visited in the second semester of 2004. They all had stoves and gas-based water heating. One hundred visits were made and they yielded 42 valid measurements. The response values range from 0.02 to 0.46, its median being 0.07. The data can be found in Zerbinatti (2008, p. 67). At the outset, we shall select the beta regression model that yields the best fit. The response is the simultaneity factor and the covariate is the release. We considered different link functions for the two sub-models (mean and precision). Model selection was based on the PRESS (prediction sum of squares) criterion; see Allen (1974). For each estimated model we computed $PRESS = \sum_{t=1}^{42} (y_t - \hat{y}_{(t)})^2 / 42$, where $\hat{y}_{(t)}$ denotes the estimate of y_t obtained after excluding such an observation from the data. The best model is the one that minimizes the criterion. The following constant dispersion model was selected: logit link and log of release used as covariate. The maximum likelihood parameter estimates are $\hat{\beta}_1 = -1.76$, $\hat{\beta}_2 = -0.76$ and $\hat{\phi} = 88.79$.

5.1 Bootstrap inference

In what follows we shall build and evaluate bootstrap-based prediction intervals for the response. To that end, we shall use the selected model, build each interval and compute

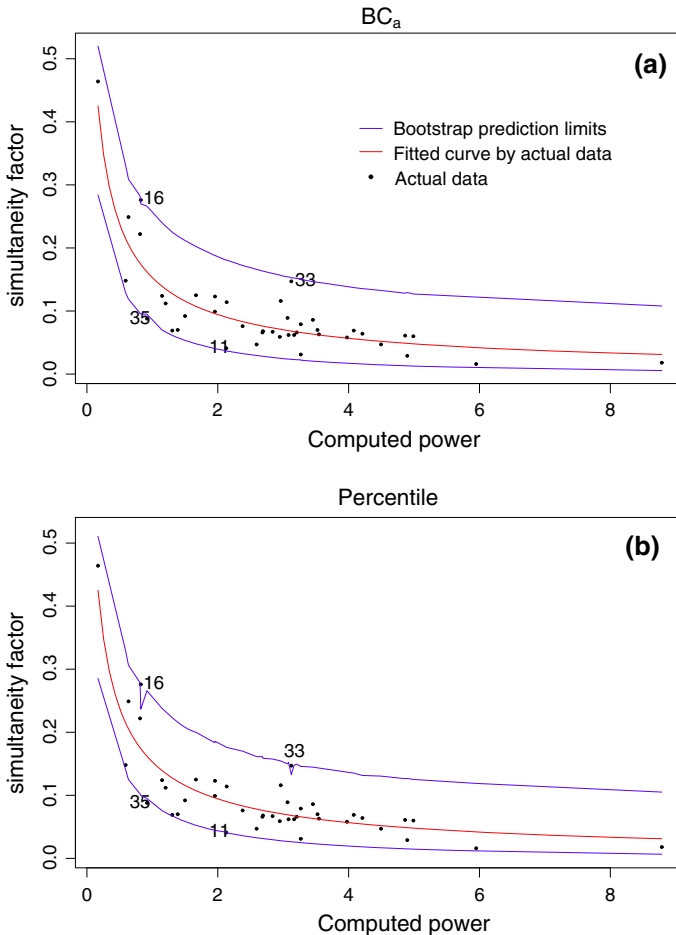


Fig. 1 Dispersion plots and prediction intervals; BC_a (a) and percentile (b)

its coverage rate using 42 different samples. Each sample is obtained by removing an observation from the data, which is done sequentially. For each sample, we construct the prediction interval and determine whether it contains the omitted response. The coverage rate is computed as the ratio between the number of prediction intervals that covered (included) the omitted response and the total number of intervals. The number of bootstrap replications was $B = 500$ and all intervals correspond to the 95% nominal level.

The empirical coverage of the percentile prediction interval was 90.4%. The intervals did not include observations 11, 16, 33 and 35. The BC_a intervals covered all omitted responses but those corresponding to observations 16 and 35, its coverage rate being 95.2%. For instance, $y_{41} = 0.041$ and the 95% BC_a interval is (0.036, 0.180). The prediction interval thus covers the response. That does not hold true for the percentile interval, which is (0.042, 0.177). Additionally, $y_{33} = 0.147$ and the BC_a and percentiles intervals are (0.023, 0.152) and (0.026, 0.147), respectively. Again, unlike

the percentile interval, the BC_a covers the omitted response. We note that the empirical coverage rates remained constant when we increased the number of bootstrap replications to $B = 2000$. The intervals lengths slightly decreased.

In Fig. 1 we plot the data (simultaneity factor vs. computed power) together with the curves created connecting the upper limits of all prediction intervals and also their lower limits. The upper panel, (a), corresponds to the BC_a interval whereas the lower panel, (b), is for percentile intervals. Figure 1 shows that the percentile prediction interval fails when it comes to observation 16; the upper interval limit bends below the observed value. The same behavior takes place when the focus lies in the prediction of y_{33} (observation 33). We also note that both intervals have similar lengths.

Acknowledgments We thank two anonymous referees for comments and suggestions and gratefully acknowledge partial financial support from CNPq and FAPESP.

References

- Allen DM (1974) The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* 16:125–127
- Cribari-Neto F, Zeileis A (2010) Beta regression in R. *J Stat Softw* 34:1–24
- Davison AC, Hinkley DV (1997) *Bootstrap methods and their application*. Cambridge University Press, New York
- DiCiccio TJ, Tibshirani R (1987) Bootstrap confidence intervals and bootstrap approximations. *J Am Stat Assoc* 82:163–170
- Efron B (1987) Better bootstrap confidence intervals. *J Am Stat Assoc* 82:171–185
- Espinheira PL, Ferrari SLP, Cribari-Neto F (2008) On beta regression residuals. *J Appl Stat* 35:407–419
- Ferrari SLP, Cribari-Neto F (2004) Beta regression for modelling rates and proportions. *J Appl Stat* 31:799–815
- Ferrari SLP, Espinheira PL, Cribari-Neto F (2011) Diagnostic tools in beta regression with varying dispersion. *Statistica Neerlandica* 65:337–351
- Mojirshiebani M, Tibshirani R (1996) Some results on bootstrap prediction intervals. *Can J Stat* 24:549–568
- Smithson M, Verkuilen J (2005) A better lemon-squeezer? Maximum likelihood regression with beta-distributed dependent variables. *Psychol Methods* 11:54–71
- Zerbinatti LFM (2008) Predição de fator de simultaneidade através de modelos de regressão para proporções contínuas, MSc thesis. University of São Paulo, Brazil. <http://www.teses.usp.br/teses/disponiveis/45/45133/tde-05042008-103844/en.php>