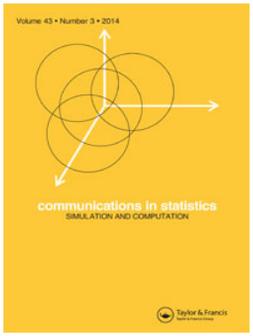

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Detecting Model Misspecification in Inflated Beta Regressions

Tarciana L. Pereira & Francisco Cribari-Neto

To cite this article: Tarciana L. Pereira & Francisco Cribari-Neto (2014) Detecting Model Misspecification in Inflated Beta Regressions, Communications in Statistics - Simulation and Computation, 43:3, 631-656, DOI: 10.1080/03610918.2012.712183

To link to this article: https://doi.org/10.1080/03610918.2012.712183

Accepted author version posted online: 16 May 2013.
Published online: 23 Sep 2013.

Communications in Statistics—Simulation and Computation®, 43: 631–656, 2014

ISSN: 0361-0918 print / 1532-4141 online
DOI: 10.1080/03610918.2012.712183

Taylor & Francis
Taylor & Francis Group

Detecting Model Misspecification in Inflated Beta Regressions

TARCIANA L. PEREIRA[1]
AND FRANCISCO CRIBARI-NETO[2]

[1]Departamento de Estatística, Universidade Federal da Paraíba, Cidade Universitária, João Pessoa/PB, Brazil
[2]Departamento de Estatística, Universidade Federal de Pernambuco, Cidade Universitária, Recife/PE, Brazil

Beta regression models are useful for modeling data that assume values in the standard unit interval (0, 1)*, such as rates and proportions. These models, however, cannot be used when the data contain observations that equal zero or one (the limits of the unit interval). Ospina and Ferrari (2012) developed the class of inflated beta regressions to handle situations in which the data contain positive mass at zero and/or one. The model is quite general and contains three submodels: for the mean, for the precision, and for the probability that the variate equals zero or one* (α)*. In this article, we propose a misspecification test for inflated beta regressions with fixed and variable dispersion. In particular, we propose two variants of the test. In the first variant, we only add testing variables to the mean submodel. The second variant follows from adding testing variables to all submodels. We perform extensive Monte Carlo simulations in order to assess the finite-sample properties of the tests (size and power), and also to gain insight on which variables should be used as testing variable and on which asymptotic testing procedure delivers the most reliable inferences. We consider a number of different misspecifications, namely: neglected nonlinearities, omitted independent variables, incorrect link functions, neglected spatial correlation, and neglected variable dispersion. Finally, an empirical illustration is presented and discussed.*

Keywords Beta distribution; Beta regression; Inflated beta regression; Link function; Misspecification test; Nonlinearity; Spatial correlation.

Mathematics Subject Classification Primary 62J02; Secondary 62J20.

1. Introduction

Oftentimes practitioners are interested in modeling the dependence of rates, proportions, and other variates that assume values in the standard unit interval on a set of explanatory variables. To that end, the beta regression model proposed by Ferrari and Cribari-Neto (2004) is particularly useful. The random variable of interest assumes values in (0, 1) and the underlying assumption is that it follows a beta law. Nonetheless, some datasets include observations on the limits of the interval, i.e., they include observations that equal

Received February 29, 2012; Accepted July 4, 2012
Address correspondence to Francisco Cribari-Neto, Departamento de Estatística, Universidade Federal da Pernambuco, Recife/PE, 50740-540, Brazil; E-mail: cribari@gmail.com; cribari@de.ufpe.br

zero and/or one. Ospina and Ferrari (2010) introduced a family of distributions known as inflated beta distributions that are mixtures of beta and Bernoulli distributions in order to allow for positive probability mass at the limits of the interval. Their distributions allow users to model data that assume values in [0, 1), (0, 1], or [0, 1]. Ospina and Ferrari (2012) introduced the class of inflated beta regression models in which the response follows a distribution that allow positive probability mass at zero or one. Their model includes a regression submodel for the probability that the dependent variable equals one of the limits of the unit interval.

When performing any regression analysis, one typically does not know whether the estimated regression model at hand provides an adequate representation for the phenomenon under study. When model specification is incorrect, inaccurate inferences are likely to follow. The development and assessment of misspecification tests is thus an important research topic in statistics and econometrics. Ramsey (1969) introduced the "Regression Specification Error Test" (RESET) for the linear regression model as a general misspecification test.

In this article, we propose two general RESET-like misspecification tests for inflated beta regressions with both fixed and variable dispersion. Misspecification means that the estimated model differs from the true data generating process in a way that the former does not provide an accurate desciption of the latter. We concentrate on two types of misspecification, namely: incorrectly specified linear predictors and incorrectly specified link functions. Our numerical (Monte Carlo) evidence shows that the proposed tests are able to detect such specification errors. In particular, we consider errors of model specification due to neglected nonlinearities, incorrectly chosen link functions, omitted covariates, neglected spatial correlation, and dispersion incorrectly taken to be constant. These are likely sources of misspecification in practical applications. It is important to note that we consider different choices of testing variables. The results indicate that the proposed tests can be quite useful for detecting model misspecification. A particular test (the simplest one) is shown to be the overall best performer in small samples.

The article unfolds as follows. The next section describes the class of inflated beta regression models with variable and fixed dispersion. The proposed misspecification tests are introduced in Section 3. In Section 4, we evaluate the finite-sample behavior of different implementations of the proposed misspecification tests using Monte Carlo simulations. An application to real data is presented in Section 5. Finally, Section 6 offers some concluding remarks.

2. Inflated Beta Regression

Ospina and Ferrari (2010) proposed distributions that are mixtures between a beta distribution and a Bernoulli distribution degenerated at 0 and/or 1 to accommodate data that are observed in (0, 1], [0, 1) and [0, 1]. The distributions are said to be inflated since they allow for positive probability mass at some points (0 and/or 1), unlike the beta law.

Under the parameterization used by Ferrari and Cribari-Neto (2004), the beta density function can be expressed as

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

where $0 < \mu < 1$ and $\phi > 0$. We say that y has beta distribution with mean μ and precision ϕ and write $y \sim \mathcal{B}(\mu, \phi)$. Here, $\mathbb{E}(y) = \mu$ and $\text{Var}(y) = \mu(1-\mu)/(1+\phi)$, where $\mathbb{E}(\cdot)$ denotes the expected value.

The cumulative distribution function of the zero- or one-inflated beta variate (inflation at a given point c; e.g., $c = 0$ or $c = 1$) is (Ospina and Ferrari, 2010)

$$\text{BI}_c(y; \alpha, \mu, \phi) = \alpha \, \mathbb{1}_{\{c\}}(y) + (1-\alpha) F(y; \mu, \phi),$$

where $\mathbb{1}_{\{c\}}(y)$ is an indicator function that equals 1 if $y = c$ and 0 otherwise, $F(\cdot; \mu, \phi)$ is the cumulative distribution function of the beta distribution $\mathcal{B}(\mu, \phi)$, and $0 < \alpha < 1$ is the mixture parameter given by $\alpha = \Pr(y = c)$.

The function BI_c has positive mass at $y = c$, and hence it is not absolutely continuous. Thus, with probability α, the variable y is selected from a distribution degenerated at c, and with probability $(1-\alpha)$ it is selected from a beta distribution.

The corresponding probability density function is (Ospina and Ferrari, 2010)

$$\text{bi}_c(y; \alpha, \mu, \phi) = \begin{cases} \alpha, & y = c, \\ (1-\alpha) f(y; \mu, \phi), & y \in (0, 1), \end{cases} \tag{1}$$

where $0 < \alpha < 1$, $0 < \mu < 1$, $\phi > 0$, and $f(y; \mu, \phi)$ is the $\mathcal{B}(\mu, \phi)$ density function. The density in Eq. (1) is inflated beta at the point c, where $c = 0$ or $c = 1$. Here, $\mathbb{E}(y) = \alpha c + (1-\alpha)\mu$ and $\text{Var}(y) = (1-\alpha)(\mu(1-\mu))/(\phi+1) + \alpha(1-\alpha)(c-\mu)^2$. If $c = 0$, the distribution in Eq. (1) is called the inflated beta distribution at zero (BEZI) and we write $y \sim \text{BEZI}(\alpha, \mu, \phi)$. If $c = 1$, it is known as the inflated beta distribution at 1 (BEOI) and we write $y \sim \text{BEOI}(\alpha, \mu, \phi)$.[1]

Ospina and Ferrari (2012) proposed inflated beta regression models, which are natural extensions of the beta regression model introduced by Ferrari and Cribari-Neto (2004). The authors assume that the response distribution is inflated beta. The continuous part of the data is modeled by the beta law and the point mass (the discrete part) is modeled using a degenerate distribution at a known value c, where c equals 0 or 1. We shall present below the class of inflated beta regression models with both fixed and variable dispersion.

Let $y_1, \ldots, y_n$ be a random sample from the inflated beta distribution at c ($c = 0$ or $c = 1$). Suppose the conditional mean, the mixture parameter, and the precision parameter satisfy the following functional relations:

$$h(\alpha_t) = \sum_{i=1}^{M} z_{ti} \gamma_i = \zeta_t, \tag{2}$$

$$g(\mu_t) = \sum_{i=1}^{m} x_{ti} \beta_i = \eta_t, \tag{3}$$

$$b(\phi_t) = \sum_{i=1}^{q} s_{ti} \lambda_i = \kappa_t, \tag{4}$$

$t = 1, \ldots, n$, where $\gamma = (\gamma_1, \ldots, \gamma_M)^\top$, $\beta = (\beta_1, \ldots, \beta_m)^\top$, and $\lambda = (\lambda_1, \ldots, \lambda_q)^\top$ are vectors of unknown regression parameters such that $\gamma \in \mathbb{R}^M$, $\beta \in \mathbb{R}^m$, and $\lambda \in \mathbb{R}^q$, ζ_t, η_t, and κ_t are linear predictors, and $x_{t1}, \ldots, x_{tm}$, $z_{t1}, \ldots, z_{tM}$, and $s_{t1}, \ldots, s_{tq}$ are fixed and

[1]For brevity, we do not consider simultaneous inflation at 0 and 1.

known covariates that may coincide totally or partly. The link functions $h : (0, 1) \to \mathbb{R}$, $g : (0, 1) \to \mathbb{R}$, and $b : (0, \infty) \to \mathbb{R}$ are strictly monotonic and twice differentiable. Several different link functions can be used, such as logit, probit, Cauchy, complementary log–log and log–log for the μ and α, and log or square root for ϕ. Here, μ_t is the mean of y_t conditional on $y_t \in (0, 1)$.

The likelihood function for $\theta = (\gamma^\top, \beta^\top, \lambda^\top)^\top$ is given by

$$L(\theta) = \prod_{t=1}^{n} \mathrm{bi}_c(y_t; \alpha_t, \mu_t, \phi_t) = L_1(\gamma) L_2(\beta, \lambda),$$

where

$$L_1(\gamma) = \prod_{t=1}^{n} \alpha_t^{\mathbb{1}_{\{c\}}(y_t)} (1 - \alpha_t)^{1 - \mathbb{1}_{\{c\}}(y_t)} \quad \text{and} \quad L_2(\beta, \lambda) = \prod_{t : y_t \in (0,1)} f(y_t; \mu_t, \phi_t),$$

where $\mathbb{1}_{\{c\}}(y)$ is the indicator function that equals 1 if $y = c$ and 0 otherwise. The parameters α_t, μ_t, and ϕ_t are functions of γ, β, and λ, respectively, through Eqs. (2), (3), and (4): $\alpha_t = h^{-1}(\zeta_t)$, $\mu_t = g^{-1}(\eta_t)$, and $\phi_t = b^{-1}(\kappa_t)$. The log-likelihood function for $\theta = (\gamma^\top, \beta^\top, \lambda^\top)^\top$ is

$$\ell(\theta) = \ell_1(\gamma) + \ell_2(\beta, \lambda),$$

where

$$\ell_1(\gamma) = \sum_{t=1}^{n} \ell_t(\alpha_t) \quad \text{and} \quad \ell_2(\beta, \lambda) = \sum_{t : y_t \in (0,1)} \ell_t(\mu_t, \phi_t),$$

with

$$\begin{aligned}
\ell_t(\alpha_t) &= \mathbb{1}_{\{c\}}(y_t) \log \alpha_t + (1 - \mathbb{1}_{\{c\}}(y_t)) \log(1 - \alpha_t), \\
\ell_t(\mu_t, \phi_t) &= \log \Gamma(\phi_t) - \log \Gamma(\mu_t \phi_t) - \log \Gamma((1 - \mu_t)\phi_t) + (\mu_t \phi_t - 1) \log y_t \\
&\quad + \{(1 - \mu_t)\phi_t - 1\} \log(1 - y_t),
\end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function.

Since the likelihood function factorizes into two terms, the parameters are separable and maximum likelihood inference on $(\beta^\top, \lambda^\top)^\top$ can be performed separately from that carried out on γ, and vice-versa. From the separability of the vectors of parameters γ and $(\beta^\top, \lambda^\top)^\top$, it is possible to independently obtain the score functions for γ and for $(\beta^\top, \lambda^\top)^\top$. Such score functions can be found in Ospina and Ferrari (2012) together with a closed-form expression for Fisher's information matrix. We note that parameter estimation is carried out by maximizing the above log-likelihood function with respect to the unknown parameters. The maximum likelihood estimators cannot be expressed in closed-form. The estimates are obtained by numerically maximizing the log-likelihood function using, for instance, a Newton or quasi-Newton algorithm.

We also note that when the precision parameter is constant for all observations, the inflated beta regression reduces to

$$h(\alpha_t) = \sum_{i=1}^{M} z_{ti} \gamma_i = \zeta_t \quad \text{and} \quad g(\mu_t) = \sum_{i=1}^{m} x_{ti} \beta_i = \eta_t,$$

$t = 1, \ldots, n$, Here, $\phi_1 = \cdots = \phi_n = \phi$, i.e., all observations share the same precision.

Finally, it is noteworthy that inflated beta regression models can be viewed as a member of the more general class of GAMLSS models ("Generalized Additive Models for Location Scale and Shape"); see Rigby and Stasinopoulos (2005). Parameter estimation can be carried using the `gamlss` package available for R (http://www.R-project.org).

We used the `GAMLSS` R package for fitting inflated beta regression models. It contains two optimization algorithms that can be used for log-likelihood maximization. The CG algorithm is a generalization of the Cole and Green (1992) algorithm and uses the first-order derivatives and the expected values of the second order and cross derivatives of the log-likelihood function with respect to parameters. Our log-likelihood maximizations were carried out using the RS algorithm, which is a generalization of the algorithm used by Rigby and Stasinopoulos (1996a, 1996b) for fitting Mean and Dispersion Additive Models (MADAM). This algorithm is well suited for situations in which the parameters are orthogonal (since the expected values of the log-likelihood cross derivatives equal zero), does not require accurate starting values for the parameters to achieve convergence (the default starting values, often constants, are usually adequate) and handles large data sets quite efficiently.

3. Testing for Model Misspecification in Inflated Beta Regressions

Ramsey (1969) introduced the "Regression Specification Error Test" (RESET) for the linear regression model as a general misspecification test. His test was tailored to detect both omitted variables and incorrect functional form. The testing strategy lies in contrasting the distribution of the residuals under correct model specification to that under the alternative hypothesis so that the model specification is in error. Under the null hypothesis of no misspecification, there will exist an efficient, consistent, and asymptotically normal estimator of the regression parameters. Under the alternative hypothesis of model misspecification, however, this estimator will be biased and inconsistent (Hausman, 1978). Ramsey and Schmidt (1976) showed that the RESET test based on least squares residuals is equivalent to the test originally proposed by Ramsey (1969). Thus, the testing procedure consists of including in the model powers of one or more variables (which are called "testing variables") and then assessing their significance using an exact F test. The underlying main idea is that when the model is correctly specified powers of the testing variables should not noticeably improve the model fit.

As noted by Godfrey (1988, p. 106), Ramsey assumes that the impact of omitted variables and other forms of model misspecification can be proxied by some (unknown) analytic function of the linear predictor used in the definition of the linear regression model: $X\beta$ (where X is the matrix of regressors and β is the vector of regression parameters). The test is then based on the underlying idea that such a function can be approximated by a polynomial. Of course, since β is a vector of unknown regression parameters it should be replaced by its ordinary least squares estimator, $\hat{\beta}$ (which equals the maximum likelihood estimator when normality is assumed). The model is augmented using powers of $X\hat{\beta}$ and the null hypothesis of no model misspecification is tested by testing the exclusion of the approximating polynomial.

Several authors evaluated the properties of the RESET test in different settings. Ramsey and Gilbert (1972) recommend that the second, third, and fourth powers of the fitted values be used as testing variables whereas Thursby and Schmidt (1977) recommend using the second, third, and fourth powers of the independent variables when augmenting the regression model. Shukur and Edgerton (2002) generalized the RESET test for systems

of simultaneous equations. Alkhamisi et al. (2008) investigated the robustness test against fat-tailed error distributions using asymptotic approximations and bootstrap resampling. Robustness against nonnormal error distributions was also investigated by Matalos and Shukur (2007). Peters (2000) used the RESET test to evaluate whether microeconometric regression models (such as the Tobit model) are misspecified. We also refer readers to Godfrey (1988, §4.2.2).

In what follows, our goal will be to determine whether an inflated beta regression is somehow misspecified. To that end, we shall propose two RESET-like misspecification tests. In the first test, we focus on the regression submodel for μ and in the second approach we devise a strategy in which all three regression submodels are augmented.

At the outset, consider the model

$$h(\alpha) = Z\gamma,$$
$$g(\mu) = X\beta,$$

where γ and β are $M \times 1$ and $m \times 1$ vectors, respectively, Z and X are $n \times M$ and $n \times m$ matrices of explanatory variables, respectively, and μ and α are $n \times 1$ vectors. Additionally, when dispersion is not constant, the precision parameter is given by

$$b(\phi) = S\lambda,$$

where λ is a $q \times 1$ vector of unknown regression parameters, S is an $n \times q$ matrix, and ϕ is an $n \times 1$ vector.

Before describing the tests, we shall introduce some notation. We write Fisher's information matrix as (Ospina and Ferrari, 2012)

$$K(\theta) = \begin{pmatrix} K_\gamma(\gamma) & 0 \\ 0 & K_\vartheta(\vartheta) \end{pmatrix},$$

where, as before, $\theta = (\gamma^\top, \beta^\top, \lambda^\top)^\top$. Here, $K_\gamma(\gamma) = K_{\gamma\gamma}$ is the Fisher information matrix for γ and

$$K_\vartheta(\beta, \lambda) = \begin{pmatrix} K_{\beta\beta} & K_{\beta\lambda} \\ K_{\lambda\beta} & K_{\lambda\lambda} \end{pmatrix}$$

is Fisher's information matrix for $\vartheta = (\beta^\top, \lambda^\top)$. The inverse of Fisher's information matrix is written, accordingly, as (Ospina and Ferrari, 2012)

$$K^{-1}(\theta) = \begin{pmatrix} K^{\gamma\gamma} & 0 & 0 \\ 0 & K^{\beta\beta} & K^{\beta\lambda} \\ 0 & K^{\lambda\beta} & K^{\lambda\lambda} \end{pmatrix}.$$

The test is carried out in two steps. First, one estimates the parameter vectors and selects a set of "testing variables," which is used to define the following "augmented model":

$$g(\mu) = X\beta + A_1\tau_1, \tag{5}$$

where A_1 is an $n \times u_1$ matrix of testing variables and τ_1 is an $u_1 \times 1$ vector of parameters. Second, we estimate Equation (5) and test the null hypothesis $\mathcal{H}_0 : \tau_1 = 0$ (the model is correctly specified) against the alternative hypothesis $\mathcal{H}_1 : \tau_1 \neq 0$ (at least one vector

element is different from zero; the model is incorrectly specified). The underlying main idea is that when the model is correctly specified such testing variables should not provide additional contribution to the quality of the model fit.

The practical implementation of the test depends upon two factors. First, one should select the testing variables to be used in the augmented regression. Second, it is necessary to select a testing procedure for the exclusion of the added variables. The testing variables used can be selected as powers of the fitted linear predictor or powers of the predicted values. The exclusion test can be performed using, e.g., the likelihood ratio, score, or Wald test.

Let $\nu = (\tau_1^\top, \gamma^\top, \beta^\top, \lambda^\top)^\top$. The likelihood ratio test statistic is

$$\xi_1 = 2\{\ell(\hat{\nu}) - \ell(\tilde{\nu})\}, \tag{6}$$

where $\hat{\nu} = (\hat{\tau_1}^\top, \hat{\gamma}^\top, \hat{\beta}^\top, \hat{\lambda}^\top)^\top$ is the (unrestricted) maximum likelihood estimators of ν and $\tilde{\nu} = (0^\top, \tilde{\gamma}^\top, \tilde{\beta}^\top, \tilde{\lambda}^\top)^\top$ is the maximum likelihood estimator of ν obtained by imposing the null hypothesis, i.e., the restricted maximum likelihood estimator. The score test statistic can be written as

$$\xi_2 = \tilde{U}_{1\beta}^\top \tilde{K}_{11}^{\beta\beta} \tilde{U}_{1\beta},$$

where $U_{1\beta}$ is the u_1-vector with the first u_1 elements of the score function $U_\beta(\beta, \lambda)$ and $K_{11}^{\beta\beta}$ is the $u_1 \times u_1$ matrix formed out of the first u_1 rows and the first u_1 columns of the inverse of $K_{\beta\beta}$. Tildes denote evaluation at the restricted maximum likelihood estimator. The Wald test statistic is

$$\xi_3 = \hat{\tau}_1^\top \big(\hat{K}_{11}^{\beta\beta}\big)^{-1} \hat{\tau}_1.$$

Here, hats denote evaluation at the unrestricted maximum likelihood estimator.

Under the usual regularity conditions and under $\mathcal{H}_0$, ξ_1, ξ_2, and ξ_3 converge in distribution to $\chi^2_{u_1}$, where u_1 is the number of elements in τ_1, a column vector. Thus, the test can be performed using approximate (asymptotic) critical values from that distribution.

We shall now introduce a second misspecification testing procedure. It follows from augmenting all three submodels, i.e., the regression submodels for μ, α, and ϕ. The augmented model is thus

$$\begin{aligned} g(\mu) &= X\beta + A_1\tau_1, \\ h(\alpha) &= Z\gamma + A_2\tau_2, \\ d(\phi) &= S\lambda + A_3\tau_3, \end{aligned}$$

where A_1, A_2, and A_3 are $n \times u_1$, $n \times u_2$, and $n \times u_3$ matrices, respectively, that contain the testing variables and τ_1, τ_2, and τ_3 are parameter vectors of dimensions $u_1 \times 1$, $u_2 \times 1$, and $u_3 \times 1$, respectively. The null hypothesis to be tested is $\mathcal{H}_0 : \tau_1 = \tau_2 = \tau_3 = 0$ (the model is correctly specified) against the alternative hypothesis that at least one τ_i, $i = 1, 2, 3$, is nonzero (the model is not correctly specified). Notice that the null hypothesis is that the model is correctly specified which is tested against the alternative that the model at hand has some form (unspecified) of model misspecification. Rejection of the null hypothesis simply means that the estimated model is poorly specified, not yielding indication of the particular form of specification error.

The likelihood ratio test statistic is as given in Equation (6) and the score and Wald test statistics can be written as sums of three quadratic forms:

$$\xi_2 = \tilde{U}_{1\gamma}^{\top}\tilde{K}_{11}^{\gamma\gamma}\tilde{U}_{1\gamma} + \tilde{U}_{1\beta}^{\top}\tilde{K}_{11}^{\beta\beta}\tilde{U}_{1\beta} + \tilde{U}_{1\lambda}^{\top}\tilde{K}_{11}^{\lambda\lambda}\tilde{U}_{1\lambda},$$
$$\xi_3 = \hat{\tau}_1^{\top}\big(\hat{K}_{11}^{\gamma\gamma}\big)^{-1}\hat{\tau}_1 + \hat{\tau}_2^{\top}\big(\hat{K}_{11}^{\beta\beta}\big)^{-1}\hat{\tau}_2 + \hat{\tau}_3^{\top}\big(\hat{K}_{11}^{\lambda\lambda}\big)^{-1}\hat{\tau}_3,$$

where $U_{1\gamma}$ is the u_2-vector with the first u_2 elements of the score function for γ, $U_\gamma(\gamma)$, and $U_{1\lambda}$ is the u_3-vector with the first u_3 elements of the score function for λ, $U_\lambda(\beta, \lambda)$. Also, $K_{11}^{\gamma\gamma}$ is defined as the $u_2 \times u_2$ matrix that contains the first u_2 rows and the first u_2 columns of $K^{\gamma\gamma}$ and $K_{11}^{\lambda\lambda}$ is defined as the $u_3 \times u_3$ matrix formed out of the first u_3 rows and the first u_3 columns of $K^{\lambda\lambda}$. Hats denote evaluation at the (unrestricted) maximum likelihood estimators and tildes denote evaluation at the restricted maximum likelihood estimator. Under the usual regularity conditions and under $\mathcal{H}_0$, ξ_1, ξ_2, and ξ_3 converge in distribution to $\chi^2_{u_1+u_2+u_3}$, where u_1, u_2, and u_3 are, respectively, the number of elements in vectors τ_1, τ_2, and τ_3. Hence, the tests can be carried out using approximate (asymptotic) critical values from that distribution.

Notice that the three tests are based on a large sample approximation to the exact null distributions of the corresponding test statistics. The test may thus display size distortions when the sample size is not large.

How reliable are the testing inferences in small samples? Which testing criterion delivers the most reliable inference in small samples? Which testing variables should be used? Is it better to only augment the mean submodel or is it better to carry out inference based on the augmentation of all three submodels? Are the tests able to identify different forms of model misspecification (e.g., incorrect links, omitted regressors, nonlinearities that have not been taken into account)? These questions shall be addressed in the next section.

4. Numerical Evaluation

We shall now present the results of a set of Monte Carlo simulations on the finite-sample behavior of the proposed misspecification tests for inflated beta regressions. The evaluating criteria are the empirical sizes (null rejection rates) and the empirical powers (nonnull rejection rates) of the tests.

The 0- or 1-inflated beta regression used in the simulations is

$$\begin{aligned} h(\alpha_t) &= \gamma_0 + \gamma_1 z_{t1}, \\ g(\mu_t) &= \beta_0 + \beta_1 x_{t1}, \\ b(\phi_t) &= \lambda_0 + \lambda_1 s_{t1}, \end{aligned} \tag{7}$$

$t = 1, \ldots, n$, where $h(\cdot)$, $g(\cdot)$, and $b(\cdot)$ are link functions and z_{t1}, x_{t1}, and s_{t1} are covariates. The covariates values were selected as follows. For each covariate, we randomly generated 40 observations from $\mathcal{U}(0, 1)$ and replicated them in order to obtain larger samples. For instance, when $n = 80$, each covariate value is replicated once (twice for $n = 120$). This is done so that the degree of nonconstant precision (measured by the ratio beween the largest and smallest precisions) is not altered when the sample size increases. All simulations were performed using `R` (see http://www.R-project.org) and were based on 10,000 replications. In each replication, we generated a random sample of the dependent variable $y = (y_1, \ldots, y_n)^{\top}$ with $y_t \sim \text{BEOI}(\alpha_t, \mu_t, \phi_t)$, where $\alpha_t = h^{-1}(\zeta_t)$, $\mu_t = g^{-1}(\eta_t)$,

and $\phi_t = b^{-1}(\kappa_t)$ (see Equations (2), (3), and (4), respectively). We use as testing variables the square of the estimated linear predictor and also the square of the predicted values; we denote the vector of predicted values by $\mu^{\dagger}$. We have also considered other testing variables (including powers of the regressors) but such results are not displayed for brevity.

Our main interest lies in evaluating the test ability to detect model misspecification. We consider several types of specification errors and, for each of them, estimate the power of the test using Monte Carlo simulations. The specification errors considered are neglected nonlinearities (i.e., the true data generating process in nonlinear, and yet a linear model is estimated), incorrect link function (i.e., the link function(s) selected are in error), omitted variables (i.e., important covariates are not included in the estimated model), neglected spatial correlation (i.e., the data display spatial correlation, and yet such a dependence is not modeled), and neglected nonconstant dispersion (i.e., the true data generating process has varying dispersion, but a fixed dispersion model is estimated).

All tests are carried out considering three nominal levels, namely: 1%, 5%, and 10%. In order for the power comparisons to be meaningful, they were performed using exact critical values estimated from the corresponding size simulations, and not asymptotic critical values. Thus, we shall compare the powers of tests that share the same size.

The test performed by only augmenting the submodel for the mean shall be denoted by R_μ and the test based on the augmentation of the three submodels shall be denoted by R_θ, with $\theta = (\mu, \alpha, \phi)^\top$ representing the vector that contains the three parameters.

Four important questions that shall be answered are: (i) Are the tests likely to detect that the regression model is incorrectly specified in small samples? (ii) Is it best to perform the misspecification test using the likelihood ratio, score, or Wald test? (iii) Which set of testing variables should be used to augment the model? (iv) Should practitioners only augment the mean submodel or should they augment all three submodels when testing for model misspecification? The numerical evidence that follows will shed some light on these questions.

4.1. *Size Simulations*

At the outset, we perform simulations under correct model specification. We test the null hypothesis of no model misspecification and the estimated model is indeed correctly specified. The response is generated according to Equation (7). For brevity, we shall only present the results of simulations performed using the logit link function for μ and α and the log link for ϕ. The true parameter values are $\gamma_0 = -2.0$, $\gamma_1 = 1.5$, $\beta_0 = -1.0$, $\beta_1 = 3.5$, $\lambda_0 = 5.1$, and $\lambda_1 = -2.8$. Table 1 contains the rejection rates (%) of the null hypothesis that the model is correctly specified. We present results for the likelihood ratio (LR), score, and Wald implementations of the R_μ test and also for the likelihood ratio implementation of the R_θ misspecification test, which was the best performing testing procedure for it.

The figures in Table 1 show that the Wald implementation of the R_μ test has the worst performance. For example, when $n = 120$ and the testing variable is $\hat{\eta}^2$, the empirical sizes of the likelihood ratio, score, and Wald tests at the 5% nominal level are, respectively, 5.58%, 5.05%, and 6.37%. Overall, the score implementation of R_μ was the best performing implementation. E.g., when $n = 80$ and the test is based on $\hat{\mu}^{\dagger 2}$, the estimated sizes of the likelihood ratio, score, and Wald tests at the 10% nominal level are 11.70%, 10.34%, and 12.84%, respectively. We also note that the best strategy is to use $\hat{\eta}^2$ as a testing variable.

We now move to the R_θ test. Recall that it follows from augmenting all three submodels. According to the notation we use, $(\hat{\eta}^2, \hat{\zeta}^2, \hat{\kappa}^2)$ indicates that the variable $\hat{\eta}^2$ is added to the regression submodel for μ, the variable $\hat{\zeta}^2$ is added to the submodel for α and the variable

Table 1
Null rejection rates (%)

R_μ test													
		$n = 40$			$n = 80$			$n = 120$			$n = 200$		
Added regressors	Test	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\hat{\eta}^2$	LR	1.88	7.55	13.44	1.3	6.15	11.45	1.26	5.58	10.81	1.21	4.96	10.22
	score	0.78	5.36	10.93	0.98	5.2	10.45	1.01	5.05	10.15	0.95	4.57	9.75
	Wald	3.45	9.82	15.86	1.98	7.15	12.69	1.57	6.37	11.40	1.35	5.22	10.64
$\hat{\mu}^{\dagger 2}$	LR	2.01	7.83	14.10	1.40	6.15	11.70	1.44	5.82	10.94	1.02	5.75	10.65
	score	0.98	5.87	11.46	1.01	5.27	10.34	1.17	5.26	10.22	0.83	5.44	10.34
	Wald	4.15	10.57	17.37	2.11	7.28	12.84	1.84	6.48	11.72	1.25	6.15	11.25
R_θ test													
		$n = 40$			$n = 80$			$n = 120$			$n = 200$		
Added regressors	Test	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$(\hat{\eta}^2, \hat{\xi}^2, \hat{\kappa}^2)$	LR	3.37	10.34	17.69	1.70	6.99	12.88	1.37	6.31	11.9	1.15	5.68	10.67
$(\hat{\eta}^2, \hat{\eta}^2, \hat{\eta}^2)$	LR	2.40	8.77	15.39	1.63	6.73	12.86	1.32	6.11	11.55	1.06	5.72	10.77
$(\hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2})$	LR	2.60	8.71	15.21	1.57	6.92	13.12	1.38	6.34	12.24	1.19	6.05	10.97
$(\hat{\eta}^2, \hat{\eta}^2, \hat{\mu}^{\dagger 2})$	LR	2.06	8.31	14.91	1.72	6.79	12.38	1.22	6.27	11.49	1.12	5.59	10.58
$(\hat{\eta}^2, \hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2})$	LR	2.14	7.86	14.43	1.64	6.66	12.47	1.23	6.15	11.66	1.10	5.63	11.01
$(\hat{\eta}^2, \hat{\mu}^{\dagger 2}, \hat{\eta}^2)$	LR	2.29	8.23	15.12	1.52	6.66	12.52	1.32	5.95	11.57	0.97	5.60	10.96

$\hat{\kappa}^2$ is added to the submodel for ϕ. Accordingly, $(\hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2})$ indicates that $\hat{\mu}^{\dagger 2}$ was used as a testing variable in all three submodels. Table 1 displays that the estimated sizes achieved with the different testing variables are similar.

Overall, the numerical evidence shows that the least size-distorted test is the score implementation of the R_μ misspecification test with $\hat{\eta}^2$ used as a testing variable. The R_θ test can be quite size-distorted and practitioners should consider applying a finite sample correction to such a test when the sample size is not large.

As suggested by a referee, we investigated the size of the score implementation of the R_μ test that uses $\hat{\eta}^2$ as testing variable when the sample size is small. We considered a sample of only 20 observations and performed simulations in which the estimated model is correctly specified. The responses were generated using the following parameter values: $\beta_0 = -0.8$, $\beta_1 = 3.5$, $\gamma_0 = -2.0$, $\gamma_1 = 1.5$, and $\phi = 70$. The null rejection rates of the likelihood ratio, score, and Wald tests at the 5% nominal level were 8.11%, 5.51%, and 10.82%, respectively. The good small sample behavior of the score test is quite noteworthy.

4.2. *Power Simulations*

Power simulations for the R_μ and R_θ tests were performed considering that only the submodel for the mean is misspecified and also under misspecification of all three submodels (except under neglected variable dispersion). In what follows, we shall only present numerical resuls for the score implementation of the R_μ test, which is the least size-distorted test. Also, we shall only report numerical results for the best and worst choices of testing variables for the R_θ test, namely $(\hat{\eta}^2, \hat{\eta}^2, \hat{\eta}^2)$ and $(\hat{\eta}^2, \hat{\zeta}^2, \hat{\kappa}^2)$ under neglected spatial correlation and $(\hat{\eta}^2, \hat{\eta}^2, \hat{\eta}^2)$ and $(\hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2})$ in all remaining cases.

4.2.1. *Nonlinearity.* Under neglected nonlinearity, the data generating process includes

$$g(\mu_t) = (\beta_0 + \beta_1 x_{t1})^\delta ,$$

with $\delta = 1.8$, $\beta_0 = 1.7$, and $\beta_1 = -1.7$. When misspecification also takes place in the other two submodels, we generate the data using

$$h(\alpha_t) = (\gamma_0 + \gamma_1 z_{t1})^\delta ,$$
$$b(\phi_t) = (\lambda_0 + \lambda_1 s_{t1})^\delta ,$$

where $\gamma_0 = 0.7$, $\gamma_1 = -0.7$, $\lambda_0 = 3.0$, and $\lambda_1 = -1.0$. Correct model specification would require $\delta = 1$. Our interest lies in assessing whether the proposed test can identify that the model specification is in error.

Table 2 presents the rejection rates under neglected nonlinearity. We note that considerably higher powers are obtained when $\hat{\eta}^2$ is used as testing variable. For instance, when $n = 80$ and at the 5% nominal level, the nonnull rejection rates of the score test are 96.85% when $\hat{\eta}^2$ is used as a testing variable and 35.39% when the testing variable is $\hat{\mu}^{\dagger 2}$. It is noteworthy that when the mean submodel is the only incorrectly specified submodel, the R_μ test with $\hat{\eta}^2$ as testing variable is more powerful than the R_θ test (in which all three submodels are augmented).

When all three submodels are incorrectly specified, the R_μ test with $\hat{\eta}^2$ as a testing variable and R_θ based on $(\hat{\eta}^2, \hat{\eta}^2, \hat{\eta}^2)$ are equally powerful. We also note that the choice of testing variable can have a decisive impact on power. As before, the power of R_μ when $\hat{\mu}^{\dagger 2}$

Table 2
Nonnull rejection rates (%): Neglected nonlinearity

Specification errors in the submodel for the mean												
R_μ test												
		$n = 40$			$n = 80$			$n = 120$			$n = 200$	
Added regressors	Tests	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\hat{\eta}^2$	Score	48.36	74.46	84.46	88.79	96.85	98.72	98.26	99.76	99.91	99.98	100.00	100.00
$\hat{\mu}^{\dagger 2}$	Score	6.87	19.22	28.14	19.08	35.39	47.06	28.52	48.69	60.48	47.24	69.19	79.29
R_θ test												
		$n = 40$			$n = 80$			$n = 120$			$n = 200$	
Added regressors	Tests	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$(\hat{\eta}^2, \hat{\eta}^2, \hat{\eta}^2)$	LR	31.89	59.39	71.34	75.63	91.45	95.61	93.69	98.76	99.59	99.87	99.97	99.99
$(\hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2})$	LR	4.63	14.16	23.20	12.84	26.85	36.84	19.46	36.57	48.18	31.34	52.70	66.16
Specification errors in the three submodels												
R_μ test												
		$n = 40$			$n = 80$			$n = 120$			$n = 200$	
Added regressors	Tests	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\hat{\eta}^2$	Score	68.69	86.79	92.09	99.16	99.80	99.91	99.98	100.00	100.00	100.00	100.00	100.00
$\hat{\mu}^{\dagger 2}$	Score	7.07	18.08	25.55	25.32	35.94	42.01	34.87	43.53	49.57	41.25	50.61	56.99
R_θ test												
		$n = 40$			$n = 80$			$n = 120$			$n = 200$	
Added regressors	Tests	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$(\hat{\eta}^2, \hat{\eta}^2, \hat{\eta}^2)$	LR	59.29	86.55	92.84	99.43	99.85	99.91	99.99	100.00	100.00	100.00	100.00	100.00
$(\hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2})$	LR	5.54	17.92	26.59	26.63	36.83	43.64	34.77	44.42	50.66	41.98	49.89	55.58

is used as a testing variable is substantially lower than when based on $\hat{\eta}^2$. Additionally, the best choice of testing variables for R_θ is $(\hat{\eta}^2, \hat{\eta}^2, \hat{\eta}^2)$.

4.2.2. *Incorrect link function.* In order to consider specification error in the link function, data are generated from Model (7) but using the complementary log–log link for μ. Here, $\beta_0 = -1.5$ and $\beta_1 = 2.5$. The estimated model incorrectly uses the logit link for μ. When all three submodels are incorrectly specified, we generate the data using the complementary log–log link for α and the inverse link for ϕ. Here, $\gamma_0 = -1.7$, $\gamma_1 = 1.0$, $\lambda_0 = 0.03$, and $\lambda_1 = -0.02$. The estimated model incorrectly uses the logit link for α and the log link for ϕ. This setup is important from a practical viewpoint since the logit link is the most commonly used link function for the mean submodel (not only in inflated beta regressions but also in beta and binary response regressions), and it would be important to have a reliable tool at hand for the detection of an incorrect link.

Table 3 presents the nonnull rejection rates (%) under misspecified link function(s). The R_μ test with $\hat{\eta}^2$ as a testing variable outperforms R_θ. Notice that the power of R_μ is considerably higher when we use $\hat{\eta}^2$ as a testing variable than when $\hat{\mu}^{\dagger 2}$ is used. The maximal power achieved by using the latter is 29.03% whereas the corresponding figure for the former is 99.95%. When all three link functions are incorrectly specified, the R_μ test with $\hat{\eta}^2$ as a testing variable again outperforms R_θ.

4.2.3. *Omitted variables.* We shall now move to the situation in which the practitioner fails to include in the regression model an important regressor. As before, we consider two scenarios. First, misspecification only takes place in the mean submodel, which is estimated using x_{t1} as the single covariate and yet the true data generating process is given by

$$g(\mu_t) = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2},$$

with $\beta_0 = -0.1$, $\beta_1 = 1.5$, and $\beta_2 = -2.0$. That is, the important independent variable x_{t2} is left out of the model. Second, when all three submodels are in error, we have, in addition, that the data generating mechanism is such that

$$h(\alpha_t) = \gamma_0 + \gamma_1 z_{t1} + \gamma_2 z_{t2},$$
$$d(\phi_t) = \lambda_0 + \lambda_1 s_{t1} + \lambda_2 s_{t2},$$

where $\gamma_0 = -0.5$, $\gamma_1 = 0.5$, $\gamma_2 = -1.5$, $\lambda_0 = 3.5$, $\lambda_1 = 1.5$, and $\lambda_2 = -3.5$. The estimated model does not include the covariates x_{t2}, z_{t2}, and s_{t2}. The results displayed in Table 4 lead to important conclusions. First, the R_μ test (only the mean submodel is augmented) is considerably more powerful when $\hat{\eta}^2$ is used as a testing variable. Second, once again R_μ is more powerful than R_θ when misspecification occurs in the mean submodel and also when all three submodels are misspecified, especially when the sample size is small. For instance, when $n = 40$ and at the 10% nominal level, the powers of R_μ and of the best performing R_θ test, when all three submodels are in error, are 59.83% and 36.83%, respectively. Third, the R_θ test performs best when $(\hat{\eta}^2, \hat{\eta}^2, \hat{\eta}^2)$ are used as testing variables. It is important to note that some of the worst performing tests (the R_μ test when $\hat{\mu}^{\dagger 2}$ is used as a testing variable and the R_θ test when $(\hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2})$ are used as testing variables) seem to be biased.

4.2.4. *Neglected spatial correlation.* Oftentimes practitioners fail to account for existing spatial correlation in their modeling strategies. It is thus important to determine whether the proposed tests can reliably detect that the model is misspecified for not taking such correlation into account. We have performed simulations in which the data are spatially correlated.

Table 3
Nonnull rejection rates (%): Incorrect link function

Specification errors in the submodel for the mean													
R_μ test													
		$n = 40$			$n = 80$			$n = 120$			$n = 200$		
Added regressors	Tests	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\hat{\eta}^2$	Score	29.81	53.61	66.23	64.79	85.48	91.95	86.81	96.24	98.34	98.83	99.85	99.95
$\hat{\mu}^{\dagger 2}$	Score	2.19	8.14	14.99	3.55	11.75	19.67	4.24	14.00	23.06	8.53	18.86	29.03
R_θ test													
		$n = 40$			$n = 80$			$n = 120$			$n = 200$		
Added regressors	Tests	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$(\hat{\eta}^2, \hat{\eta}^2, \hat{\eta}^2)$	LR	17.87	37.58	51.11	46.87	70.88	81.07	70.50	89.08	94.07	95.18	98.88	99.54
$(\hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2})$	LR	2.22	8.61	15.86	3.33	12.79	20.54	5.0	15.09	24.92	8.22	21.97	33.59
Specification errors in the three submodels													
R_μ test													
		$n = 40$			$n = 80$			$n = 120$			$n = 200$		
Added regressors	Tests	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\hat{\eta}^2$	Score	22.14	45.73	57.82	48.78	74.81	84.11	75.10	90.45	94.55	95.58	98.68	99.33
$\hat{\mu}^{\dagger 2}$	Score	1.86	7.41	13.36	2.28	8.05	14.30	2.83	9.13	15.75	3.26	10.33	16.93
R_θ test													
		$n = 40$			$n = 80$			$n = 120$			$n = 200$		
Added regressors	Tests	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$(\hat{\eta}^2, \hat{\eta}^2, \hat{\eta}^2)$	LR	10.44	29.16	40.76	29.72	55.94	68.34	55.80	77.34	85.94	86.52	95.06	97.74
$(\hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2})$	LR	1.93	8.05	15.40	3.35	12.12	19.49	4.81	14.42	22.93	7.02	19.55	29.95

Table 4
Nonnull rejection rates (%): Omitted variables

Specification errors in the submodel for the mean													
R_μ test													
		$n = 40$			$n = 80$			$n = 120$			$n = 200$		
Added regressors	Tests	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\hat{\eta}^2$	Score	59.87	80.03	89.77	94.14	98.76	99.55	99.52	99.96	99.98	99.99	100.00	100.00
$\hat{\mu}^{\dagger 2}$	Score	0.89	3.77	8.23	1.17	4.45	8.60	1.10	4.06	8.75	1.04	4.03	8.71
R_θ test													
		$n = 40$			$n = 80$			$n = 120$			$n = 200$		
Added regressors	Tests	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$(\hat{\eta}^2, \hat{\eta}^2, \hat{\eta}^2)$	LR	38.34	66.95	79.14	84.58	95.62	97.92	97.97	99.60	99.88	99.98	99.99	100.00
$(\hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2})$	LR	0.99	4.63	9.61	1.33	4.84	9.60	1.34	4.90	9.09	1.15	4.78	9.51
Specification errors in the three submodels													
R_μ test													
		$n = 40$			$n = 80$			$n = 120$			$n = 200$		
Added regressors	Tests	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\hat{\eta}^2$	Score	23.78	46.16	59.83	55.13	79.47	86.88	80.04	93.12	96.56	97.16	99.17	99.67
$\hat{\mu}^{\dagger 2}$	Score	1.17	5.30	10.50	1.69	5.97	11.20	1.76	6.33	11.68	2.28	7.97	13.73
R_θ test													
		$n = 40$			$n = 80$			$n = 120$			$n = 200$		
Added regressors	Tests	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$(\hat{\eta}^2, \hat{\eta}^2, \hat{\eta}^2)$	LR	7.54	24.04	36.83	29.05	53.14	66.15	51.15	72.88	82.35	81.82	93.71	96.56
$(\hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2}, \hat{\mu}^{\dagger 2})$	LR	0.72	3.94	8.71	1.40	6.03	10.96	1.52	6.46	12.07	2.08	8.18	14.11

To that end, we generated latent Gaussian spatially correlated effects and included them in the model linear predictor. The model is estimated without taking into account the spatial correlation. The mean submodel parameter values are $\beta_0 = -0.8$ and $\beta_1 = 3.5$. When all three submodels are in error of specification, we have, in addition, that $\gamma_0 = -2.0$ and $\gamma_1 = 1.5$ (submodel for α) and $\lambda_0 = 6.0$ and $\lambda_1 = -3.5$ (submodel for ϕ).

The simulation results are presented in Table 5. Important conclusions can be drawn from these results. First, the powers of all tests are low when the sample is small. Large samples are needed to reliably detect neglected spatial correlation. That is why we now report results for sample sizes that range from 40 to 1,000 observations. Second, once again the R_μ test was considerably more powerful when the mean submodel was augmented using $\hat{\eta}^2$. When $n = 200$ and at the 5% nominal level, the power of the score implementation of R_μ based on $\hat{\eta}^2$ is 40.68% whereas the corresponding power when the inference is based on $\hat{\mu}^{\dagger 2}$ is 6.62%. Third, unlike the previous misspecification scenarios, R_μ is now less powerful than R_θ. For instance, when $n = 200$ and at the 10% nominal level, the powers of the best performing R_μ and R_θ tests are, respectively, 62.22% and 96.31% when misspecification only takes place in the mean submodel, and 35.06% and 96.23% under misspecification in all three submodels. Nonetheless, we notice that R_μ has good power when the sample size is large. Again, it is important to note that some of the worst performing tests (the R_μ test when $\hat{\mu}^{\dagger 2}$ is used as testing variable and the R_θ test when $(\hat{\eta}^2, \hat{\zeta}^2, \hat{\kappa}^2)$ are used as testing variables) seem to be biased.

4.2.5. *Nonconstant dispersion.* An important property of a misspecification test lies in its ability to identify the lack of fit due to nonconstant dispersion when the model is estimated under the incorrect assumption that dispersion is fixed, i.e., that ϕ is constant across observations. We have performed simulations to assess whether our test is able to detect such a lack of fit. Here, we only consider inference based on the R_μ test. The responses were generated using the following parameter values: $\beta_0 = -0.8$, $\beta_1 = 3.5$, $\gamma_0 = -2.0$, $\gamma_1 = 1.5$, $\lambda_0 = 6.0$, and $\lambda_1 = -6.0$. Thus, the variability in ϕ was generated through a regression structure for ϕ; see Equation (7). Parameter estimation was carried out assuming that ϕ is constant, an incorrect assumption. The results presented in Table 6 show that the score implementation delivers the most powerful test. The results also show that the test based on $\hat{\eta}^2$ is once again considerably more powerful than that based on $\hat{\mu}^{\dagger 2}$.

We also ran simulations for different values of $\lambda_1 : -1, -2, -3, \ldots$. The estimated power curve corresponding to the 5% nominal level and $n = 200$ for the score implementation of the R_μ test is presented in Fig. 1. Power curves illustrate the effect on power of varying the alternative hypothesis. Note that the test is quite powerful in correctly rejecting the null hypothesis when the true value of λ_1 is less than -5.

4.3. *Size and Power Simulations Under Constant Dispersion*

A simulation study, similar to that performed for the inflated beta regression with variable dispersion, was carried out for inflated beta regression models with constant dispersion. The constant dispersion-inflated beta regression used in the simulations was

$$h(\alpha_t) = \gamma_0 + \gamma_1 z_{t1},$$
$$g(\mu_t) = \beta_0 + \beta_1 x_{t1},$$

$t = 1, \ldots, n$, where $h(\cdot)$ and $g(\cdot)$ are link functions and z_{t1} and x_{t1} are covariates. In all simulations, $\phi = 70$. We randomly generated 40 observations of each covariate from

Table 5
Nonnull rejection rates (%): Neglected spatial correlation

Specification errors in the submodel for the mean													
R_μ test													
		$n = 40$		$n = 80$		$n = 120$		$n = 200$		$n = 400$		$n = 1000$	
Added regressors	Tests	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%
$\hat{\eta}^2$	Score	2.79	7.25	8.58	19.54	16.97	32.79	40.68	62.22	85.22	93.91	99.98	100.00
$\hat{\mu}^{\dagger 2}$	Score	1.37	3.45	2.02	5.91	2.93	8.46	6.62	15.38	20.09	34.44	68.30	82.89
R_θ test													
		$n = 40$		$n = 80$		$n = 120$		$n = 200$		$n = 400$		$n = 1000$	
Added regressors	Tests	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%
$(\hat{\eta}^2, \hat{\eta}^2, \hat{\eta}^2)$	LR	11.51	21.39	32.37	48.74	57.56	73.69	89.84	96.31	99.96	100.00	100.00	100.00
$(\hat{\eta}^2, \hat{\xi}^2, \hat{\kappa}^2)$	LR	3.90	9.12	14.04	25.27	26.34	41.47	57.24	73.52	95.66	98.51	100.00	100.00
Specification errors in the three submodels													
R_μ test													
		$n = 40$		$n = 80$		$n = 120$		$n = 200$		$n = 400$		$n = 1000$	
Added regressors	Tests	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%
$\hat{\eta}^2$	Score	1.45	4.68	4.18	10.67	7.87	18.22	19.04	35.06	51.47	69.53	96.46	99.01
$\hat{\mu}^{\dagger 2}$	Score	2.87	6.52	1.16	3.42	1.20	3.96	2.38	5.93	5.06	10.23	16.16	29.30
R_θ test													
		$n = 40$		$n = 80$		$n = 120$		$n = 200$		$n = 400$		$n = 1000$	
Added regressors	Tests	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%
$(\hat{\eta}^2, \hat{\eta}^2, \hat{\eta}^2)$	LR	16.18	26.33	39.95	55.08	63.84	77.27	91.22	96.23	99.96	99.99	100.00	100.00
$(\hat{\eta}^2, \hat{\xi}^2, \hat{\kappa}^2)$	LR	1.65	4.91	9.51	18.17	16.04	27.10	32.90	48.29	73.82	84.58	99.69	99.85

Table 6
Nonnull rejection rates (%): Nonconstant dispersion

		R_μ test											
		$n = 40$			$n = 80$			$n = 120$			$n = 200$		
Added regressors	Tests	1%	5%	10%	1%	5%	10%	1%	5%	10%	1%	5%	10%
$\hat{\eta}^2$	LR	39.00	59.18	69.20	69.76	84.31	89.26	86.65	94.09	96.41	97.69	99.27	99.66
	Score	42.07	61.01	70.01	71.67	84.49	89.50	87.95	94.30	96.47	97.71	99.29	99.65
	Wald	32.49	56.10	68.08	66.77	83.70	88.97	85.63	94.04	96.33	97.51	99.30	99.66
$\hat{\mu}^{\dagger 2}$	LR	17.93	34.34	44.88	39.61	58.60	68.41	57.33	73.48	82.29	78.38	91.48	95.0
	Score	20.34	34.32	44.62	40.24	58.33	67.93	56.58	73.57	82.02	78.52	91.20	94.82
	Wald	14.06	32.46	43.98	36.79	57.54	68.01	56.12	73.0	82.06	77.92	91.39	95.02

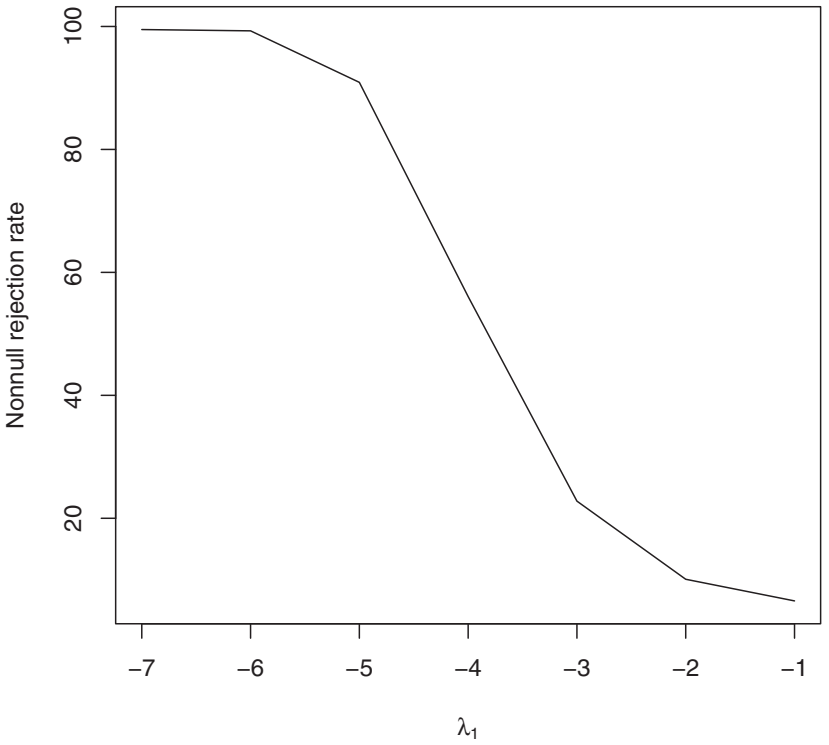

Figure 1. Estimated power curve.

the standard uniform distribution and then replicated them twice, three, and five times in order to obtain covariate values to be used when the sample size is 80, 120, and 200, respectively. The reported results are based on 10, 000 Monte Carlo replications. The specification errors considered in the power simulations are neglected nonlinearity in the linear predictor, omitted regressor, incorrect choice of link function, and neglected spatial correlation.

In the size simulations, the response values are generated using logit links, $\beta_0 = -1.0$, $\beta_1 = 3.5$, $\gamma_0 = -1.5$, and $\gamma_1 = 1.0$. Under unaccounted nonlinearity, data generation used

$$g(\mu_t) = (\beta_0 + \beta_1 x_{t1})^{\delta},$$

with $\delta = 1.8$, $\beta_0 = 1.9$, and $\beta_1 = -1.9$. When misspecification takes place in both the submodels (for μ and α), we have, additionally, that

$$h(\alpha_t) = (\gamma_0 + \gamma_1 z_{t1})^{\delta},$$

where $\gamma_0 = 0.7$ and $\gamma_1 = -0.7$. Here, we shall use the following notation: $\theta' = (\mu, \alpha)^{\top}$. Hence, $R_{\theta'}$ denotes the test in which we augment the two submodels (μ and α).

Under incorrect choice of link function, we generate data using the complementary log–log link function for μ and estimate the model assuming that the link is logit. Here, $\beta_0 = -1.5$ and $\beta_1 = 2.5$. When, in addition, the submodel for α is incorrectly specified, data generation uses the complementary log–log link and estimation is based on the logit link for that submodel as well. In that case, $\gamma_0 = -1.7$ and $\gamma_1 = 1.0$.

Under omitted variable, the data generating process includes a second regressor, x_{t2}, in the mean submodel. The parameter values are $\beta_0 = -0.1$, $\beta_1 = 1.5$, and $\beta_2 = -2.0$. The estimated model does not include x_{t2}. When misspecification reaches the submodel for α, data generation includes an additional regressor, z_{t2}, which is not included in the estimated model. Here, $\gamma_0 = -0.5$, $\gamma_1 = 0.5$, and $\gamma_2 = -1.5$.

Finally, we generate data with spatial correlation and estimate the regression parameters without accounting for such a correlation. Here, $\beta_0 = -1.0$ and $\beta_1 = 3.5$. When the submodel for α is also in error, we use $\gamma_0 = -1.5$ and $\gamma_1 = 1.0$.

For brevity, we only report results for the score implementation of R_μ test using $\hat{\eta}^2$ as a testing variable and for the likelihood ratio implementation of $R_{\theta'}$ based on $(\hat{\eta}^2, \hat{\eta}^2)$, i.e., when $\hat{\eta}^2$ is used to augment the submodel for μ and the submodel for α; these are the best performing tests. The Monte Carlo simulation results are presented in Table 7. We note that the size distortions of the two tests are small, with slight advantage for R_μ. We also conclude that the tests are satisfactorily powerful, except under neglected spatial correlation, in which case the sample size must be large in order for the type II error probability to be small. Interestingly and unlike what we observed under variable dispersion, R_μ is more powerful than $R_{\theta'}$ when one fails to account for spatial correlation. (Recall that under variable dispersion the test we called R_θ is the equivalent of $R_{\theta'}$ here.) Indeed, under both misspecifications only in the mean submodel and misspecifications in the two submodels (μ and α), R_μ typically outperforms $R_{\theta'}$.

5. An Empirical Application

We shall now apply the proposed test to a real (not simulated) dataset. The response values are efficiency indices for municipalities in the state of São Paulo, Brazil. The data are a subset of that used by Sampaio de Souza et al. (2005) in the sense that we focus on the state of São Paulo (the richest state in Brazil) whereas the authors considered all municipalities in the country. Another difference is that we use a variable which was not used by Sampaio de Souza et al. (2005), namely: the municipality age (equal to 1 if less than 8 years and 0 otherwise). Several municipalities were created after 1988 in Brazil, mostly by spliting an existing one into two, and it is important to evaluate the impact of that on how well counties are run.

It is important to point out that the data contain 1's (which correspond to fully efficient units) and that we use the inflated beta regression whereas the authors used linear and quantile regressions. The data we use contain 427 municipalities of the state of São Paulo for the year 2000. Tables 8 and 9 present a brief description of the variables along with some descriptive statistics. The variables E6, E7, E8, E9, E10, E11, E12, and E13 are indicators that identify the political party to which the mayor is affiliated. Note that all variables in Table 9 are dummies, i.e., they only assume values 0 and 1. The regressor W1EFIC (not listed in the tables) is obtained as the product between a neighborhood matrix (diagonal elements equal to 0, off-diagonal elements equal to 1 if the distance between the two counties does not exceed 50 km and 0 otherwise) and the vector of responses; it is included in the regression to account for existing spatial correlation.

At the outset, we select and estimate an inflated (at 1) beta regression model. The response is the set of efficiencies that were computed for the state of São Paulo. The mean efficiency equals 0.66 (standard deviation: 0.19), and there are 32 units that are fully efficient (i.e., the data contain 32 observations that are equal to 1—about 7.5% of all observations). Our goal is to jointly model the mean efficiency, the data dispersion, and the probability of a given unit being fully efficient using the inflated beta regression framework. Table 10

Table 7
Null and nonnull rejection rates (%) under fixed dispersion

Size									
			$n = 40$		$n = 80$		$n = 120$		$n = 200$	
Scenario	Tests	Added regresors	5%	10%	5%	10%	5%	10%	5%	10%
R_μ test	Score	$\hat{\eta}^2$	5.68	11.07	5.37	10.57	5.16	10.08	4.77	9.97
$R_{\theta'}$ test	LR	$(\hat{\eta}^2, \hat{\eta}^2)$	6.96	13.31	5.70	11.01	5.45	11.01	5.40	10.49
Power: Nonlinearity										
			$n = 40$		$n = 80$		$n = 120$		$n = 200$	
Scenario	Tests	Added regressors	5%	10%	5%	10%	5%	10%	5%	10%
R_μ test with error in μ	Score	$\hat{\eta}^2$	89.19	94.19	99.58	99.90	99.99	100.00	100.00	100.00
$R_{\theta'}$ test with error in μ	LR	$(\hat{\eta}^2, \hat{\eta}^2)$	82.90	90.51	99.25	99.64	99.98	99.99	100.00	100.00
R_μ test with error in μ and α	Score	$\hat{\eta}^2$	65.56	76.68	93.28	96.70	99.16	99.65	99.98	99.99
$R_{\theta'}$ test with error in μ and α	LR	$(\hat{\eta}^2, \hat{\eta}^2)$	58.86	71.40	90.32	94.89	98.49	99.35	99.98	99.99
Power: Incorrect link function										
			$n = 40$		$n = 80$		$n = 120$		$n = 200$	
Scenario	Tests	Added regressors	5%	10%	5%	10%	5%	10%	5%	10%
R_μ test with error in μ	Score	$\hat{\eta}^2$	58.00	70.63	88.06	93.46	97.30	98.92	99.83	99.97
$R_{\theta'}$ test with error in μ	LR	$(\hat{\eta}^2, \hat{\eta}^2)$	46.70	60.41	81.83	89.14	94.50	97.34	99.58	99.81
R_μ test with error in μ and α	Score	$\hat{\eta}^2$	58.24	70.83	88.33	93.34	97.38	99.00	99.85	99.95
$R_{\theta'}$ test with error in μ and α	LR	$(\hat{\eta}^2, \hat{\eta}^2)$	47.16	60.76	82.19	89.24	84.87	94.61	99.51	99.81

(Continued on next page)

Table 7
Null and nonnull rejection rates (%) under fixed dispersion *(Continued)*

Power: Omitted variable										
			$n = 40$		$n = 80$		$n = 120$		$n = 200$	
Scenario	Tests	Added regressors	5%	10%	5%	10%	5%	10%	5%	10%
R_μ test with error in μ	Score	$\hat{\eta}^2$	86.85	92.89	99.43	99.78	99.96	99.99	100.00	100.00
$R_{\theta'}$ test with error in μ	LR	$(\hat{\eta}^2, \hat{\eta}^2)$	79.27	88.73	98.54	99.50	99.90	99.96	100.00	100.00
R_μ test with error in μ and α	Score	$\hat{\eta}^2$	83.18	90.56	98.93	99.50	99.95	99.98	100.00	100.00
$R_{\theta'}$ test with error in μ and α	LR	$(\hat{\eta}^2, \hat{\eta}^2)$	75.07	85.22	97.81	99.21	99.84	99.95	100.00	100.00
Power: Neglected spatial correlation										
			$n = 40$		$n = 80$		$n = 120$		$n = 200$	
Scenario	Tests	Added regressors	5%	10%	5%	10%	5%	10%	5%	10%
R_μ test with error in μ	Score	$\hat{\eta}^2$	9.08	19.39	26.64	41.93	44.75	63.45	75.78	88.35
$R_{\theta'}$ test with error in μ	LR	$(\hat{\eta}^2, \hat{\eta}^2)$	6.19	13.24	17.53	30.90	31.59	48.15	60.98	77.32
R_μ test with error in μ and α	Score	$\hat{\eta}^2$	8.32	19.30	28.86	46.94	51.51	71.24	84.55	93.13
$R_{\theta'}$ test with error in μ and α	LR	$\hat{\eta}^2, (\hat{\eta}^2)$	6.63	13.73	22.24	37.41	39.70	57.68	74.48	87.45

Table 8
Continuous variables

Variable	Description	Mean (standard error)
EFIC	Efficiency scores DEA-CCR	0.66 (0.19)
DESP	Personnel expenses	1.95e+07 (9.32e+07)
SAL	% heads of households that earn up to 1 minimum wage	2.01 (1.84)
REND	Average earnings	700.15 (232.37)
CAD	Updatedness of the real state register	0.91 (0.12)
DENS	Demographic density	306.89 (1146.85)
URB	Urbanization rate	82.45 (15.67)

displays the maximum likelihood estimates of the parameters that index the inflated beta model. The model was chosen after an extensive model selection process in which many candidate models were considered. The logit link is used for μ and α and the log link is used for ϕ.

Our next goal is to test whether the estimated model is correctly specified. The score R_μ test based on $\hat{\eta}^2$ and the likelihood ratio R_θ test based on $(\hat{\eta}^2, \hat{\eta}^2, \hat{\eta}^2)$ were used. The score R_μ test statistic equals 4.94 and the likelihood ratio R_θ test statistic equals 6.70. Therefore, the correct model specification is not rejected at the 1% signficance level when we use R_μ and at the 5% nominal level when inference is based on R_θ.

We removed the W1EFFIC regressor (which accounts for spatial correlation) from the mean and precision submodels, estimated the model without such a covariate, and performed the specification test R_μ using $\hat{\eta}^2$ as testing variable. The test was carried out using its score

Table 9
Dummy variables

Variable	Description	Frequency of ones (%)
E6	PFL	52 (12.18)
E7	PMDB	76 (17.80)
E8	PSDB	118 (27.63)
E9	PT	30 (7.02)
E10	PPS	24 (5.62)
E11	PPB	21 (4.92)
E12	PTB	50 (11.71)
E13	PDT	14 (3.28)
E17	Participation in intermunicipal consortia	75 (17.56)
E18	Degree of computer utilization	390 (91.33)
E19	Decision power of municipal councils	217 (50.82)
E22	Alvorada Project	3 (0.70)
E25	Municipality age	36 (8.43)
MT	Tourist destination	72 (16.86)
R2	Royalties	55 (12.88)

Table 10
Parameter estimates of the variable dispersion-inflated beta using the efficiency scores dataset

Submodel for μ		
Variable	Estimate	Standard error
Constant	0.1403	0.2012
W1EFIC	0.0142	0.0080
E4	−0.0007	0.0001
E9	0.2776	0.1495
E21	0.0095	0.0026
Submodel for α		
Variable	Estimate	Standard error
Constant	7.774e-01	1.271e+00
E1	9.363e-09	4.915e-09
E3	−5.244 e-01	2.267e-01
E4	−2.835 e-03	1.511e-03
E17	1.357e+00	4.296e-01
E18	−1.068 e+00	5.248e-01
Submodel for ϕ		
Variable	Estimate	Standard error
Constant	2.0711	0.3028
W1EFIC	−0.0384	0.0139
E4	0.0012	0.0003
E18	−0.7416	0.2538

test implementation. The value of the test statistic was 5.20 with corresponding p-value 0.0226. We thus reject the null hypothesis that the estimated model is correctly specified at the 5% nominal level. Additionally, we have also estimated a fixed-dispersion-inflated beta regression. The R_μ test statistic equals 5.07, the p-value being 0.0243. Again, the null hypothesis of no model misspecification is rejected at the 5% nominal level. Therefore, in both cases (no regressor that accounts for spatial correlation and fixed dispersion), the test points to incorrect model formulation.

6. Concluding Remarks

In this article, we proposed a general misspecification test for inflated beta regressions with both fixed and variable dispersion. In particular, we proposed two variants of the test. In the first variant, we only augment the mean submodel whereas in the second one all three submodels are augmented with testing variables. The null hypothesis that the model is correctly specified is rejected whenever the added variables noticeably improve the quality of the model fit. The proposed tests are based on a large sample approximation

and can be carried out using the likelihood ratio, score, or Wald criteria. Several important questions were addressed in an extensive numerical evaluation: (i) How reliable are the testing inferences in small samples? (ii) Which testing criterion delivers the most reliable inference? (iii) Which testing variables should be used? (iv) Is it better to only augment the mean submodel or is it better to carry out inference based on the augmentation of all three submodels? (v) Are the tests able to identify different forms of model misspecification (e.g., incorrect links, omitted regressors, nonlinearities that have not been taken into account)?

We considered different sources of model misspecification that are likely to take place in empirical analyses: neglected nonlinearity, incorrect choice of link function, omitted regressors, neglected variable dispersion, and neglected spatial correlation. The results showed that the proposed test has good power in small to moderate sample sizes, except when the practitioner fails to account for spatial correlation, in which case large sample sizes are needed to achieve small type II error probabilities. The numerical evidence has also shown that the simplest of the two tests, i.e., that in which only the mean submodel is augmented, is typically the best performing one. Our results also reveal that the choice of testing variables has a sizeable impact on the test finite-sample performance. Practitioners should be careful when choosing such variables. Our strong recommendation is that model augmentation be based on the squared values of the fitted linear predictor.

We believe that the proposed test can be quite useful in practical situations and we recommend to practitioners to model data using inflated beta regressions and use it to assess whether their models are correctly specified.

Acknowledgments

TLP and FCN gratefully acknowledge financial support from CAPES and CNPq, respectively. The authors also thank three anonymous referees for comments and suggestions.

References

Alkhamisi, M. A., Khalaf, G., Shukur, G. (2008). The effect of fat-tailed error terms on the properties of systemwise RESET test. *Journal of Applied Statistics* 35:101–113.

Cole, T. J., Green, P. J. (1992). Smoothing reference centile curves: The LMS method and penalized likelihood. *Statistics in Medicine* 11:1305–1319.

Ferrari, S. L. P., Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics* 31:799–815.

Godfrey, L. G. (1988). *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches*. CambridgeMA:Cambridge University Press.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica* 46:1251–1271.

Matalos, P., Shukur, G. (2007). The robustness of the RESET test to non-normal error terms. *Computational Economics* 30:393–408.

Ospina, R., Ferrari, S. L. P. (2010). Inflated beta distributions. *Statistical Papers* 51:111–126.

Ospina, R., Ferrari, S. L. P. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics and Data Analysis* 56:1609–1623.

Peters, S. (2000). On the use of the RESET test in micro-econometric models. *Applied Economics Letters* 7:361–365.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society B* 31:350–371.

Ramsey, J. B., Gilbert, R. (1972). A Monte Carlo study of some small sample properties of tests for specification error. *Journal of the American Statistical Association* 67:180–186.

Ramsey, J. B., Schmidt, P. (1976). Some further results on the use of OLS and BLUS residuals in specification error tests. *Journal of the American Statistical Association* 71:389–390.

Rigby, R. A., Stasinopoulos, D. M. (1996a). A semi-parametric additive model for variance heterogeneity. *Statistics and Computing* 6:57–65.

Rigby, R. A., Stasinopoulos, D. M. (1996b). Mean and dispersion additive models. In: Hardle, W., Schimek, M. G., eds., *Statistical Theory and Computational Aspects of Smoothing*. Heidelberg: Physica-Verlag, pp. 215–230.

Rigby, R. A., Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Applied Statistics* 54:507–554.

Sampaio de Souza, M. C., Cribari-Neto, F., Stosic, B. D. (2005). Explaining DEA technical efficiency scores in an outlier corrected environment: The case of public services in Brazilian municipalities. *Brazilian Review of Econometrics* 25:289–315.

Shukur, G., Edgerton, D. L. (2002). The small sample properties of the RESET test as applied to systems of equations. *Journal of Statistical Computation and Simulation* 72:909–924.

Thursby, J. G., Schmidt, P. (1977). Some properties of tests for specification error in a linear regression model. *Journal of the American Statistical Association* 72:635–641.