machine learning & knowledge extraction

MDPI

Article

Model Selection Criteria on Beta Regression for Machine Learning

Patrícia L. Espinheira, Luana C. Meireles da Silva, Alisson de Oliveira Silva and Raydonal Ospina *

Departamento de Estatística, CAST – Computational Agriculture Statistics Laboratory, Universidade Federal de Pernambuco, Recife 50740-540, Brazil; patespipa@de.ufpe.br (P.L.E.); lcmds1@de.ufpe.br (L.C.M.d.S.); ados1@de.ufpe.br or allysson_jlr@yahoo.com.br (A.d.O.S.)

* Correspondence: raydonal@de.ufpe.br or raydonal@castlab.org

Received: 19 January 2019; Accepted: 6 February 2019; Published: 8 February 2019

Abstract: Beta regression models are a class of supervised learning tools for regression problems with univariate and limited response. Current fitting procedures for beta regression require variable selection based on (potentially problematic) information criteria. We propose model selection criteria that take into account the leverage, residuals, and influence of the observations, both to systematic linear and nonlinear components. To that end, we propose a Predictive Residual Sum of Squares (PRESS)-like machine learning tool and a prediction coefficient, namely P^2 statistic, as a computational procedure. Monte Carlo simulation results on the finite sample behavior of prediction-based model selection criteria P^2 are provided. We also evaluated two versions of the R^2 criterion. Finally, applications to real data are presented. The new criterion proved to be crucial to choose models taking into account the robustness of the maximum likelihood estimation procedure in the presence of influential cases.

Keywords: beta regression; influence; residuals; PRESS; P^2 criterion; R^2-like criteria

1. Introduction

The class of nonlinear beta regression models was proposed by [1] and extended to situations in which the data include zeros and/or ones by [2,3]. Shortly thereafter, [4] developed for the class of nonlinear beta regression model residuals and measures of local influence. Local influence proposed by [5] is a decisive scheme to select the model that fit well a dataset, takes into account that the estimation process is robust in influential cases. Indeed, the final conclusion about model selection should consider the analysis of influence. The model selection is a crucial step in data analysis, since all inference performance is based on the selected model. [6] evaluated the behavior of different model selection criteria in a beta regression model, such as the Akaike Information Criterion (AIC) [7], Schwarz Bayesian Criterion (SBC) [8] and various approaches based on pseudo-R^2.

However, it is common for models selected by the usual selection criteria to present poorly fitted or influential observations. Indeed, the best models selected by the usual criteria do not always present residual plots that validate the goodness-of-fit. Also, current fitting procedures for beta regression are infeasible for high-dimensional data setups and require variable selection based on (potentially problematic) information criteria. Furthermore, the usual selection criteria do not offer any insight about the quality of the predictive values. In this context, [9] proposed the PRESS (Predictive Residual Sum of Squares) statistic, which can be used as a measure of the predictive power of a model. [10] proposed a coefficient of prediction based on PRESS, namely P^2 that is similar to the R^2 approach. The P^2 statistic can

be used to select models from a predictive perspective [11]. Moreover, the PRESS statistic presents a close relationship with the Cook distance [12] and local influence measures [5], as we shall present here. Hence, the P^2 statistic is a selection criterion that takes into account the impact of the observations poorly fitted by the model, observations with atypical residuals, and cases that exert a disproportional effect on the model estimation process, even affect the inference conclusions, influential cases.

Our main goal is to introduce a PRESS-like machine learning tool and its associated P^2 statistic, as a coefficient of prediction for the class of nonlinear beta regression models. The P^2 statistic is used as a selection criterion for beta regression. We carried out Monte Carlo simulations to evaluate the behavior of the P^2 measure, as well as the behavior of two usual R^2-like criterion, namely: the R^2_{LR} [6] and R^2_{FC} [13]. We considered a variety of simulation scenarios, as different ranges for the response mean, several sample sizes and values to the precision parameter, five link functions and different model misspecifications. Finally, we present and discuss two applications to real data.

The simulation and application data set results showed that small values of the new criterion are an indication that the robustness of the maximum likelihood estimation procedure of the model in the presence of influential points is worthy of further investigation. This information could not be accessed by usual selection criteria. However, the issue about variability is still better accessed by R^2-like criteria. Thus, the best machine learning strategy is to use the three criteria discussed here to choose the best model, once each one holds on different information.

2. P^2 Criterion

Consider the linear model $Y = X\beta + \varepsilon$ as the supervised learning procedure, where Y is a vector $n \times 1$ of the responses, X is a known matrix of covariates (measured features) of dimension $n \times p$ of full rank, β is the parameter vector of dimension $p \times 1$ and ε is a vector $n \times 1$ of errors. We have the least-squares estimators: $\widehat{\beta} = (X^\top X)^{-1} X^\top y$, the residual $r_t = y_t - x_t^\top \widehat{\beta}$ and the predicted value $\widehat{y}_t = x_t^\top \widehat{\beta}$, where $x_t^\top = (x_{t1}, \ldots, x_{tp})$, and $t = 1, \ldots, n$. Notice that we found these quantities in one shot, without doing any sort of iterative optimization. Let $\widehat{\beta}_{(t)}$ be the least-squares estimate of β without the tth observation and $\widehat{y}_{(t)} = x_t^\top \widehat{\beta}_{(t)}$ be the predicted value of the case deleted, such that $r_{(t)} = y_t - \widehat{y}_{(t)}$ is the prediction error. Thus, for multiple regression, the classic Predictive Residual Sum of Squares statistic named here as PRESS_C is given by

$$\text{PRESS}_C = \sum_{t=1}^{n} r^2_{(t)} = \sum_{t=1}^{n} (y_t - \widehat{y}_{(t)})^2 = \left(\frac{r_t}{1 - h_{tt}} \right)^2, \tag{1}$$

where r_t is the ordinary residual obtained by regressing y on X and h_{tt} is the tth diagonal element of the projection matrix $X(X^\top X)^{-1} X^\top$ of this regression.

Now, let $y_1, \ldots, y_n$ be independent random variables, such that each y_t, for $t = 1, \ldots, n$, is beta distributed denoted by $y_t \sim \mathcal{B}(\mu_t, \phi_t)$, i.e., each y_t has density function given by

$$f(y_t; \mu_t, \phi_t) = \frac{\Gamma(\phi_t)}{\Gamma(\mu_t \phi_t)\Gamma((1-\mu_t)\phi_t)} y_t^{\mu_t \phi_t - 1} (1 - y_t)^{(1-\mu_t)\phi_t - 1}, \quad 0 < y_t < 1, \tag{2}$$

where $0 < \mu_t < 1$ and $\phi_t > 0$. Here, $\text{E}(y_t) = \mu_t$ and $\text{Var}(y_t) = V(\mu_t)/(1 + \phi_t)$, where $V(\mu_t) = \mu_t(1 - \mu_t)$. Ref. [1] proposed the class of nonlinear beta regression models in which the mean of y_t and the precision parameter can be written as

$$g(\mu_t) = \eta_{1t} = f_1(x_t^\top; \beta) \quad \text{and} \quad h(\phi_t) = \eta_{2t} = f_2(z_t^\top, \gamma), \quad t = 1, \ldots, n, \tag{3}$$

where $\beta = (\beta_1, \ldots, \beta_k)^\top$ and $\gamma = (\gamma_1, \ldots, \gamma_q)^\top$ are, respectively, $k \times 1$ and $q \times 1$ vectors of unknown parameters ($\beta \in \mathbb{R}^k$; $\gamma \in \mathbb{R}^q$), η_{1t} and η_{2t} are the nonlinear predictors, $x_t^\top = (x_{t1}, \ldots, x_{tk_1})$ and $z_t^\top =$

$(z_{t1}, \ldots, z_{tq_1})$ are vectors of covariates (i.e., vectors of known variables), $t = 1, \ldots, n$, $k_1 \leq k$, $q_1 \leq q$ and $k + q < n$. Both $g(\cdot)$ and $h(\cdot)$ are strictly monotonic and twice differentiable link functions. Furthermore, $f_i(\cdot)$, $i = 1, 2$, are differentiable and continuous functions, such that the matrices $J_1 = \partial\eta_1/\partial\beta$ and $J_2 = \partial\eta_2/\partial\gamma$ have full rank (their ranks are equal to k and q, respectively). The model's parameters can be estimated by maximum likelihood (ML). In the Appendix, we present the log-likelihood function, the score vector and Fisher's information matrix for the nonlinear beta regression model. Model (2)–(3) embodies the beta regression linear model with varying dispersion when the predictors are linear functions of the parameters. In this case, $g(\mu_t) = \eta_{1t} = x_t^\top \beta$ and $h(\phi_t) = z_t^\top \gamma$. If, in addition, the predictor for μ_t is linear and ϕ_t is constant through the observations, we arrive at the beta regression model defined [13].

The beta regression likelihood is inherently nonlinear and there are no closed form expressions for the ML estimators, and their computations should be performed numerically using a nonlinear optimization algorithm for machine learning, e.g., some form of iterative Newton's method (Newton–Raphson, Fisher's scoring, BHHH, etc.). To propose a PRESS-like statistic for beta regression, we shall explore the relationship between the Fisher iterative ML scheme and a weighted least square regression. This regression considers pseudo variables as proposed [14] to build a Cook-like distance that have been used in several classes of regression models. Fisher's scoring iterative scheme used for estimating β, both to linear and nonlinear regression model, can be written as

$$\beta^{(m+1)} = \beta^{(m)} + (K_{\beta\beta}^{(m)})^{-1} U_\beta^{(m)}(\beta), \tag{4}$$

where $m = 0, 1, 2, \ldots$ are the iterations which are carried out until convergence. The convergence happens when the difference $|\beta^{(m+1)} - \beta^{(m)}|$ is less than a small, previously specified constant.

From Appendix's expressions (A1), (A2) and from (4), it follows that the mth scoring iteration for β, in a class of nonlinear regression model is defined as

$$\beta^{(m+1)} = \beta^{(m)} + (J_1^\top \Phi W J_1)^{-1} J_1^\top \Phi T(y^* - \mu^*), \tag{5}$$

where the tth elements of the vectors y^* and μ^* are

$$y_t^* = \log\{y_t/(1 - y_t)\} \quad \text{and} \quad \mu_t^* = \psi(\mu_t\phi_t) - \psi((1 - \mu_t)\phi_t), \quad t = 1, \ldots, n, \tag{6}$$

$\Phi = \text{diag}(\phi_1, \ldots, \phi_n)$, $T = \text{diag}(1/g'(\mu_1), \ldots, 1/g'(\mu_n))$ $J_1 = \partial\eta_1/\partial\beta$, $W = \text{diag}(w_1, \ldots, w_n)$, w_t given in Appendix, expression (A3). Furthermore, Φ, W, J_1, T and μ^* are evaluated at β^m.

Ref. [14] suggests that we rewrite the iterative process in (5) by defining the following vector $u_1^{(m)} = J_1\beta^{(m)} + W^{-1}T(y^* - \mu^*)$ such that the equation in (5) becomes

$$\beta^{(m+1)} = (J_1^\top \Phi W J_1)^{-1} \Phi J_1^\top W u_1^{(m)}. \tag{7}$$

Upon convergence, we may write the ML estimator of β as

$$\widehat{\beta} = (J_1^\top \widehat{\Phi}\widehat{W} J_1)^{-1} \widehat{\Phi} J_1^\top \widehat{W} u_1, \quad \text{where} \quad u_1 = J_1\widehat{\beta} + \widehat{W}^{-1}\widehat{T}(y^* - \widehat{\mu}^*). \tag{8}$$

Here, $\widehat{W}$, $\widehat{\Phi}$, $\widehat{T}$ and $\widehat{J}_1$ are the matrices W, Φ, T and J_1, respectively, evaluated at the ML estimators of β and γ. We note that $\widehat{\beta}$ in (8) can be viewed as the least-squares estimator of β obtained by regressing the pseudo observation vector

$$y^\dagger = \widehat{\Phi}^{1/2}\widehat{W}^{1/2} u_1 \text{ on } J_1{}^\dagger = \widehat{\Phi}^{1/2}\widehat{W}^{1/2} J_1. \tag{9}$$

Since we had the expression of $\widehat{\beta}$ in (8), several quantities related to the pseudo regression in (9) may be obtained, as the ordinary residual, the projection matrix, the $\widehat{\beta}_{(t)}$ and the prediction error. Following [14] we have that

$$\widehat{\beta}_{(t)} = \widehat{\beta} - \{(J_1^\top \widehat{\Phi}\widehat{W}J_1)^{-1} J_{1t}\widehat{\phi}_t^{1/2}\widehat{w}_t^{1/2} r_t^\beta\}/(1-h_{tt}^*), \tag{10}$$

in which $J_{1t}^\top$ is the tth row of the J_1 matrix, h_{tt}^* and r_t^β, are respectively, the tth diagonal element of the projection matrix $H^* = (\widehat{W}\widehat{\Phi})^{1/2} J_1 (J_1\widehat{\Phi}\widehat{W}J_1)^{-1} J_1^\top (\widehat{\Phi}\widehat{W})^{1/2}$ and the tth ordinary residual of the pseudo regression in (8) given by

$$r_t^\beta = \widehat{\phi}_t^{1/2}\widehat{w}_t^{1/2} u_{1,t} - \widehat{\phi}_t^{1/2}\widehat{w}_t^{1/2} J_{1t}^\top \widehat{\beta} = \frac{y_t^* - \widehat{\mu}_t^*}{\sqrt{\widehat{v}_t}}, \tag{11}$$

where $u_{1,t}$ the tth element of the vector u_1 and v_t are given in (8) and (A3–Appendix), respectively. We note that $\mu^* = \mathrm{E}(y^*)$, $v_t = \mathrm{Var}(y_t^*)$ and r_t^β in (11) is the *standardized weighted residual 1* [15]. In what follows we able to define $\widehat{y}_{(t)}^\dagger = \widehat{\phi}_t^{1/2}\widehat{w}_t^{1/2} J_{1t}^\top \widehat{\beta}_{(t)}$ and the prediction error

$$r_{(t)}^\dagger = y_t^\dagger - \widehat{y}_{(t)}^\dagger = \widehat{\phi}_t^{1/2}\widehat{w}_t^{1/2} u_{1,t} - \widehat{\phi}_t^{1/2}\widehat{w}_t^{1/2} J_{1t}^\top \widehat{\beta}_{(t)}. \tag{12}$$

Plugging (10) quantity in (12) we then obtain that $r_{(t)}^\dagger$ is

$$\begin{aligned} y_t^\dagger - \widehat{y}_{(t)}^\dagger &= \widehat{\phi}_t^{1/2}\widehat{w}_t^{1/2} u_{1,t} - \widehat{\phi}_t^{1/2}\widehat{w}_t^{1/2} J_{1t}^\top \left\{ \widehat{\beta} - \frac{\{(J_1^\top \widehat{\Phi}\widehat{W}J_1)^{-1} J_{1t}\widehat{\phi}_t^{1/2}\widehat{w}_t^{1/2} r_t^\beta\}}{(1-h_{tt}^*)} \right\} \\ &= \widehat{\phi}_t^{1/2}\widehat{w}_t^{1/2} u_{1,t} - \widehat{\phi}_t^{1/2}\widehat{w}_t^{1/2} J_{1t}^\top \widehat{\beta} - \frac{\{\widehat{\phi}_t^{1/2}\widehat{w}_t^{1/2} J_{1t}^\top (J_1^\top \widehat{\Phi}\widehat{W}J_1)^{-1} J_{1t}\widehat{\phi}_t^{1/2}\widehat{w}_t^{1/2} r_t^\beta\}}{(1-h_{tt}^*)} \\ &= r_t^\beta - h_{tt}^* r_t^\beta/(1-h_{tt}^*) = r_t^\beta/(1-h_{tt}^*). \end{aligned}$$

Finally, for the nonlinear beta regression models, the PRESS-like statistic becomes as

$$\mathrm{PRESS} = \sum_{t=1}^{n} (y_t^\dagger - \widehat{y}_{(t)}^\dagger)^2 = \sum_{t=1}^{n} \left(\frac{r_t^\beta}{1-h_{tt}^*} \right)^2. \tag{13}$$

It is noteworthy that based on (11) the expression in (13) may be written as

$$\mathrm{PRESS} = \sum_{t=1}^{n} \frac{(r_{p,t}^\beta)^2}{1-h_{tt}^*}, \tag{14}$$

in which

$$r_{p,t}^\beta = \frac{y_t^* - \widehat{\mu}_t^*}{\sqrt{\widehat{v}_t(1-h_{tt}^*)}}. \tag{15}$$

It is important to emphasize that $r_{p,t}^\beta$ in the PRESS expression given in (14) is the *standardized weighted residual 2* proposed by [15] which outperforming the others beta regression residuals in their numerical evaluation. In special, outperforming the ordinary residual $(y_t - \widehat{\mu}_t)/\sqrt{\mathrm{Var}(y_t)}$. The weighted residuals are constructed using the difference between the logit of the responses and their fitted means, the main qualities of beta distribution. The same holds to the proposal for the PRESS-like statistic to the class of beta regression.

Indeed, the PRESS also has relationships with influence measures [16]. Ref. [12] use a version of the likelihood displacement [17] to build the Cook's distance that measures the impact of a given observation on the parameter estimates of the mean sub-model by removing it from the data. Based on approximations for the likelihood displacement, the Cook-like distances have been proposed for several classes of regression models. Focus on beta regression models [18] obtain an approximation for the version of likelihood displacement to build the Cook's distance given by

$$LD_t = \frac{(r_{p,t}^{\beta})^2}{1-h_{tt}^*} h_{tt}^* \Rightarrow \frac{(r_{p,t}^{\beta})^2}{1-h_{tt}^*} = \frac{LD_t}{h_{tt}^*}, \tag{16}$$

plugging (16) measure in (13) we thus obtain that

$$\text{PRESS} = \sum_{t=1}^{n} \frac{LD_t}{h_{tt}^*}, \tag{17}$$

in which LD_t is the Cook-like distance to the class of beta regression. Furthermore, Ref. [19] shown that $LD_t \approx C_t$ in which C_t is the total local influence of observation t, defined in (A4). Thus, PRESS $\approx \sum_{t=1}^{n} \frac{C_t}{h_{tt}^*}$, which clarify the relationship of this statistic with the local influence measures. Ref. [19] also suggest that observations such that $C_t > 2\sum_{t=1}^{n} C_t/n$ can be taken to be individually influential. We shall take the same threshold pattern to highlight influential observations on the index plots of Cook-like distances. When the predictors in (3) are linear functions of the parameters, i.e., $g(\mu_t) = x_t^\top \beta$ and $h(\phi_t) = z_t^\top \gamma$, the expression in (17) also represents the PRESS-like statistic for the class of the linear beta regression models with $p = k + q$ unknown regression parameters.

Considering the same approach to build the determination coefficient R^2 in linear models, Ref. [10] define a prediction measure based on the PRESS_C statistic in (1) as $P_C^2 = 1 - \text{PRESS}_C/SST_{(t)}$, where $SST_{(t)} = \sum_{t=1}^{n}(y_{(t)} - \overline{y}_{(t)})^2$, with $\overline{y}_{(t)}$ being the arithmetic mean for $n-1$ values of y_t.

Based on this idea, for beta regression models we must use the pseudo regression procedure defined in (9) to build the P^2-like coefficient as

$$P^2 = 1 - \frac{\text{PRESS}}{SST_{(t)}^{\dagger}}, \tag{18}$$

where $SST_{(t)}^{\dagger} = \sum_{t=1}^{n}(y^{\dagger} - \overline{y}_{(t)}^{\dagger})^2$, with $\overline{y}_{(t)}^{\dagger}$ being the arithmetic average for $n-1$ values of the vector $y^{\dagger} = \widehat{\Phi}^{1/2}\widehat{W}^{1/2}u_1$ by excluding the tth observation. It can be shown that $SST_{(t)}^{\dagger} = (n/(n-p))^2 SST^{\dagger}$, wherein p is the number of model parameters and SST is the Total Sum of Squares for the full data. For the class of nonlinear beta regression models, $SST^{\dagger} = \sum_{t=1}^{n}(y_t^{\dagger} - \overline{y}^{\dagger})^2$, where $\overline{y}^{\dagger}$ is the arithmetic average of the $y_t^{\dagger}$, $t = 1, \ldots, n$ and $p = k + q$. Please note that P^2 given in (18) is not a positive quantity. Indeed, the $\text{PRESS}/SST^{\dagger}{}_{(t)}$ is a positive quantity, thus the P^2 take values in $(-\infty; 1]$. The closer to one, the better is the predictive power of the model.

To compare the behaviors of the P^2 defined in (18) and R^2-like criteria we consider at the outset two versions of pseudo-R^2 based on the likelihood ratio. The first one was proposed by [20] as $R_{LR}^2 = 1 - (L_{null}/L_{fit})^{2/n}$, where L_{null} is the ML achievable (saturated model) and L_{fit} is the likelihood achieved by the model under investigation. The second version is a proposal of [6] that takes into account the inclusion of covariates both in the mean and in the precision sub-models, is given by: $R_{LR_c}^2 = 1 - (1 - R_{LR}^2)\left(\frac{n-1}{n-(1+\alpha)k_1-(1-\alpha)q_1}\right)^{\delta}$, where $\alpha \in [0,1]$ and $\delta > 0$. Based on simulation presented by the authors we chose $\alpha = 0.4$ and $\delta = 1$. We also consider the R_{FC}^2, which is defined as the square of the sample coefficient of correlation between $g(y)$ and $\widehat{\eta}_1$ [13], and its penalized version based on [6] given by $R_{FC_c}^2 = 1 - (1 - R_{FC}^2)(n-1)/(n-(k_1+q_1))$, where k_1 and q_1 are, respectively, the number of covariates

of the mean and dispersion sub-models. By analogy, we define the penalized version of P^2 given by $P_c^2 = 1 - (1 - P^2)(n-1)/(n-(k_1+q_1))$.

3. Simulation Study

The Monte Carlo experiments present in this section were carried out using both fixed and varying dispersion beta regressions as data generating processes, as well as linear and nonlinear models. All simulations were carried out using the `Ox` matrix programming language [21]. The number of Monte Carlo replications is 10,000. Our goal is simultaneously to assess the performance of the P^2, R^2_{FC} and R^2_{LR} criteria, and, additionally, which values, on average, these statistics could assume under different data settings and features of the regression model. To that end, at the outset, we present the average values of the statistics as the arithmetic mean of the Monte Carlo replicas. Also, we provide information about the distributions of the statistics by a boxplot analysis.

Since the upper limits of all statistics are equal to one, a performance evaluation criterion for these measures is that their values go to one if the model is correctly specified and far from one otherwise. The mean values of the statistics are especially useful when the scenarios considered in the simulations occur in the real data analysis.

3.1. Linear Setting: Fixed Dispersion, Omitted Covariates and Link Functions

Table 1 shows the mean values of the statistics obtained by simulation of the constant dispersion beta regression model that involves a systematic component for the mean given by

$$\log\left(\frac{\mu_t}{1-\mu_t}\right) = \beta_1 + \beta_2\, x_{t2} + \beta_3\, x_{t3} + \beta_4\, x_{t4} + \beta_5\, x_{t5}, \quad t = 1, \ldots, n, \tag{19}$$

that is based on logit link function. The covariate values were independently obtained as random draws of the following distributions: $X_{ti} \sim U(0,1)$, $i = 2, \ldots, 5$ and were kept fixed throughout the experiment. The precisions, the sample sizes and the range of mean response are, respectively, $\phi = (20, 50, 150, 400, 1000)$, $n = (40, 80, 120, 400)$, $\mu \in (0.005, 0.12)$, $\mu \in (0.90, 0.99)$ and $\mu \in (0.20, 0.88)$. Under the model specification given in (19) we investigate the behavior of the statistics by omitting covariates. In this case, we considered the Scenarios 1, 2, and 3, in which are omitted three, two, and one covariate, respectively. In a fourth scenario, the estimated model is correctly specified.

The results in Table 1 show that the mean values of all statistics increase as important covariates are included in the model and the value of ϕ increases. On the other hand, as the size of the sample increases, the model misspecification is evidenced by lower values of the statistics (Scenarios 1, 2, and 3). It shall be noted that the mean values for all statistics are considerably larger when $\mu \in (0.20, 0.88)$. Additionally, their values approach one when the estimated model is closest to the true model. For instance, in Scenario 4 for $n = 40$, $\phi = 150$ the values of P^2 and R^2_{LR} are, respectively, 0.936 and 0.947.

The behavior of the statistics for finite sample size changes substantially when $\mu \in (0.90; 0.99)$. It is noteworthy the reduction of its mean values, in special to the P^2 criterion when $\mu \approx 1$ revealing the difficulty in fitting the model in this range of μ. Even under true specification (Scenario 4) the P^2 criterion identifies unmistakably some problem in the model-fitting when $\mu \approx 1$. For instance, when $n = 80$ and $\phi = 50$, we have $P_c^2 = -0.007$ and $R^2_{LR_c} = 0.542$. The same feature occurs when $\mu \in (0.005, 0.12)$.

Table 1. Mean values of the statistics. True model versus misspecification models (omitted covariates (Scenarios 1, 2, and 3)). The model estimated correctly: Scenario 4.

		Scenario 1			**Scenario 2**			**Scenario 3**			**Scenario 4**		
	Estimated model	$g(\mu_t)=\beta_1+\beta_2\,x_{t2}$			$g(\mu_t)=\beta_1+\beta_2\,x_{t2}+\beta_3\,x_{t3}$			$g(\mu_t)=\beta_1+\beta_2\,x_{t2}+\beta_3\,x_{t3}+\beta_4\,x_{t4}$			$g(\mu_t)=\beta_1+\beta_2\,x_{t2}+\beta_3\,x_{t3}+\beta_4\,x_{t4}+\beta_5\,x_{t5}$		
		$\mu\in(0.20,0.88)$; $\beta=(-1.9,1.2,1.0,1.1,1.3)^\top$.											
n	$\phi\to$	20	50	150	20	50	150	20	50	150	20	50	150
40	P^2	0.307	0.363	0.393	0.393	0.463	0.502	0.506	0.602	0.656	0.694	0.847	0.936
	P_c^2	0.270	0.329	0.361	0.342	0.418	0.461	0.450	0.557	0.617	0.649	0.825	0.927
	R^2_{LR}	0.296	0.358	0.391	0.394	0.473	0.515	0.518	0.620	0.675	0.723	0.869	0.947
	$R^2_{LR_c}$	0.258	0.324	0.358	0.344	0.429	0.475	0.463	0.577	0.638	0.682	0.849	0.939
80	P^2	0.286	0.346	0.379	0.368	0.445	0.488	0.506	0.584	0.643	0.666	0.833	0.930
	P_c^2	0.267	0.329	0.363	0.343	0.423	0.468	0.450	0.561	0.624	0.643	0.821	0.925
	R^2_{LR}	0.291	0.356	0.391	0.385	0.468	0.513	0.518	0.614	0.672	0.706	0.860	0.943
	$R^2_{LR_c}$	0.273	0.339	0.375	0.361	0.447	0.494	0.463	0.593	0.655	0.686	0.851	0.939
		$\mu\in(0.90,0.99)$; $\beta=(1.8,1.2,1,1.1,0.9)^\top$.											
n	$\phi\to$	20	50	150	20	50	150	20	50	150	20	50	150
40	P^2	0.119	0.061	0.071	0.139	0.062	0.072	0.171	0.072	0.156	0.149	0.089	0.213
	P_c^2	0.071	0.010	0.021	0.067	−0.016	−0.006	0.076	−0.034	0.059	0.023	−0.045	0.097
	R^2_{LR}	0.164	0.196	0.243	0.221	0.266	0.336	0.271	0.374	0.466	0.444	0.593	0.774
	$R^2_{LR_c}$	0.119	0.153	0.203	0.157	0.205	0.281	0.188	0.303	0.405	0.362	0.533	0.741
80	P^2	0.093	0.036	0.044	0.112	0.038	0.046	0.149	0.045	0.120	0.123	0.056	0.175
	P_c^2	0.070	0.011	0.019	0.077	0.000	0.008	0.103	−0.006	0.073	0.063	−0.007	0.119
	R^2_{LR}	0.158	0.190	0.240	0.211	0.253	0.327	0.268	0.356	0.451	0.416	0.571	0.760
	$R^2_{LR_c}$	0.136	0.169	0.221	0.180	0.224	0.301	0.229	0.321	0.422	0.376	0.542	0.744
		$\mu\in(0.005,0.12)$; $\beta=(-1.5,-1.2,-1.0,-1.1,-1.3)^\top$.											
n	$\phi\to$	20	50	150	20	50	150	20	50	150	20	50	150
40	P^2	0.128	0.063	0.056	0.108	0.059	0.028	0.153	0.070	0.202	0.149	0.090	0.212
	P_c^2	0.081	0.013	0.005	0.033	−0.020	−0.053	0.056	−0.036	0.111	0.023	−0.044	0.096
	R^2_{LR}	0.199	0.215	0.254	0.265	0.349	0.379	0.326	0.415	0.548	0.442	0.595	0.774
	$R^2_{LR_c}$	0.156	0.172	0.214	0.204	0.295	0.327	0.249	0.348	0.496	0.360	0.535	0.741
80	P^2	0.105	0.040	0.032	0.083	0.043	0.012	0.128	0.038	0.165	0.123	0.057	0.174
	P_c^2	0.081	0.015	0.006	0.047	0.005	−0.027	0.081	−0.013	0.121	0.064	−0.007	0.119
	R^2_{LR}	0.197	0.211	0.251	0.253	0.340	0.372	0.311	0.394	0.534	0.416	0.572	0.760
	$R^2_{LR_c}$	0.176	0.191	0.231	0.223	0.314	0.347	0.274	0.362	0.509	0.376	0.543	0.743

In what follows, we shall investigate the empirical distributions of the statistics: P^2, P_c^2, R^2_{LR}, $R^2_{LR_c}$, R^2_{FC} and $R^2_{FC_c}$ under the correctly specified modeling (scenario 4) in Table 1, for $n=40$ and $\phi=150$. These results are shown using boxplots of 10,000 values of the statistics obtained from the Monte Carlo simulations (Figure 1). The mean value of the statistic replications is represented by a dot on the side of each boxplot. In panels (a), (b) and (c) we present the boxplots for $\mu\approx 0$, μ scattered on the standard unit interval and for $\mu\approx 0$, respectively.

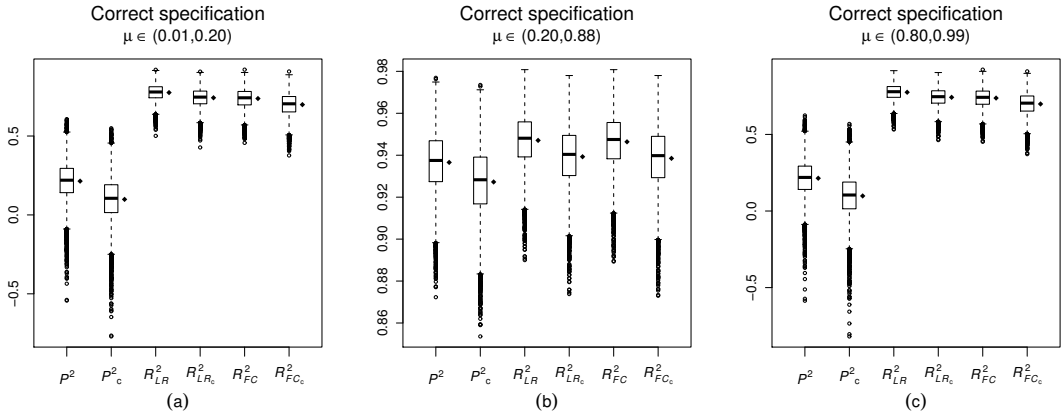

Figure 1. Correct specification: $g(\mu_t) = \beta_1 + \beta_2\, x_{t2} + \beta_3\, x_{t3} + \beta_4\, x_{t4} + \beta_5\, x_{t5}$, $n = 40$, $\phi = 150$. (**a**) $\mu \approx 0$; (**b**) μ scattered on the standard unit interval; (**c**) $\mu \approx 1$.

Figure 1 shows that the means and medians of all statistics are close, thus the mean values of the statistics adequately represent their behavior in these scenarios. We also notice that both P^2 and R^2 criteria are so small, for models correctly specified when μ is close to the boundaries of the standard unit interval (Figure 1). However, it is noteworthy how the P^2 values are substantially smaller than the R^2-like criterion values.

When the mean response is concentrated on the boundaries of the standard unit interval, even when the model is correctly specified, the statistic of prediction assumes negative values, panel (a) and panel (c). Based on panel (b) ($\mu \in (0.20, 0.88)$), it can be seen that when the response mean response is scattered on the standard unit interval, the behavior of the prediction statistic is very different, with values much more concentrated nearby one. The same behavior occurs for the goodness-of-fit measures. R^2_{FC} and R^2_{LR}.

In Figure 2 we consider a misspecification problem (three omitted covariates). For illustration, we consider only $\phi = 50$ and $n = 40, 80, 120, 400$, $\mu \in (0.20, 0.88)$. We notice that when three covariates are omitted, with the increasing of sample size, the replication values of the statistics tend to concentrate at small values, as expected due to the misspecification problem.

We notice that typically the mean and the median of the 10,000 values of the statistic is closed, confirming the usefulness of the mean values to describe these measures. When $n = 400$ (panel d), the values of all statistics tend to concentrate around a value far from 1, i.e., around 0.3, and 0.4. It behaves noteworthy as the prediction and determination coefficients behave equally in this scenario ($\mu \in (0.20, 0.88)$).

In Figure 3 we consider $\mu \in (0.01, 0.20)$ and the model is estimated correctly, $n = 40$ and $\phi = (50, 150, 400, 1000)$. We notice that the values of the R^2-like statistics become more concentrated and closer to one as the value of ϕ increases. Nonetheless, the behavior of P^2 statistics is quite different. Even when $\phi = 400$ this measure displays negative values (panel(c)). These observations that present P^2 negative values are cases, poorly fitted by the model and potential influential cases. It is noteworthy that cases poorly fitted by the model can befall in despite of that $\phi = 1000$ (Figure 3d). The statistics present the same feature when $\mu \approx 1$.

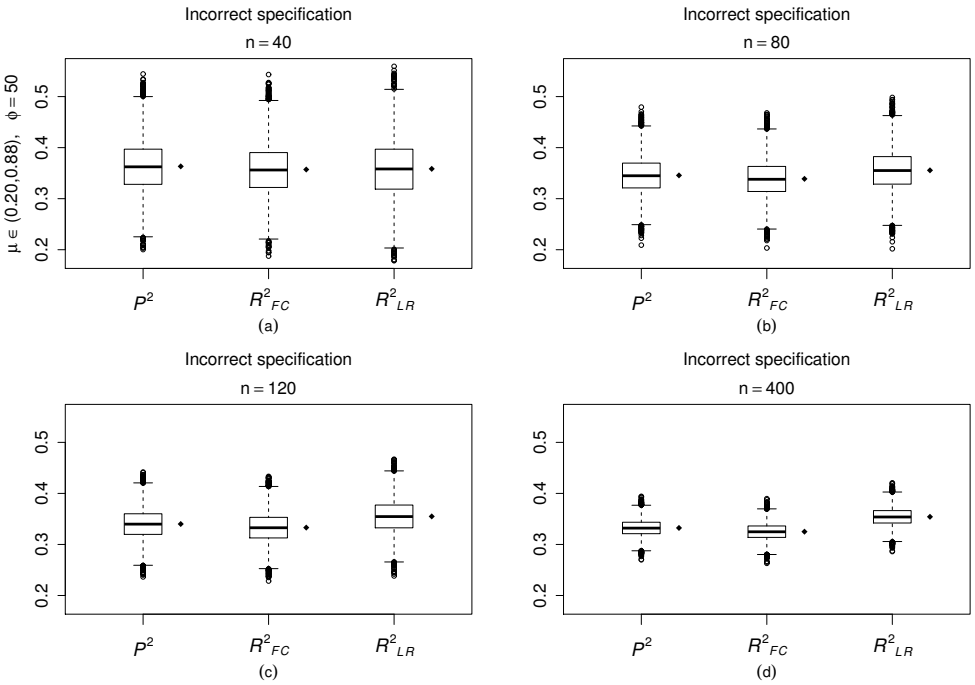

Figure 2. Omitted covariates. Estimated model: $g(\mu_t) = \beta_1 + \beta_2\, x_{t2}$. Correct model: $g(\mu_t) = \beta_1 + \beta_2\, x_{t2} + \beta_3\, x_{t3} + \beta_4\, x_{t4} + \beta_5\, x_{t5}$; $\mu \in (0.20, 0.88)$; $\phi = 50$. (**a**) sample size $n = 40$; (**b**) sample size $n = 80$; (**c**) sample size $n = 120$; (**d**) sample size $n = 200$.

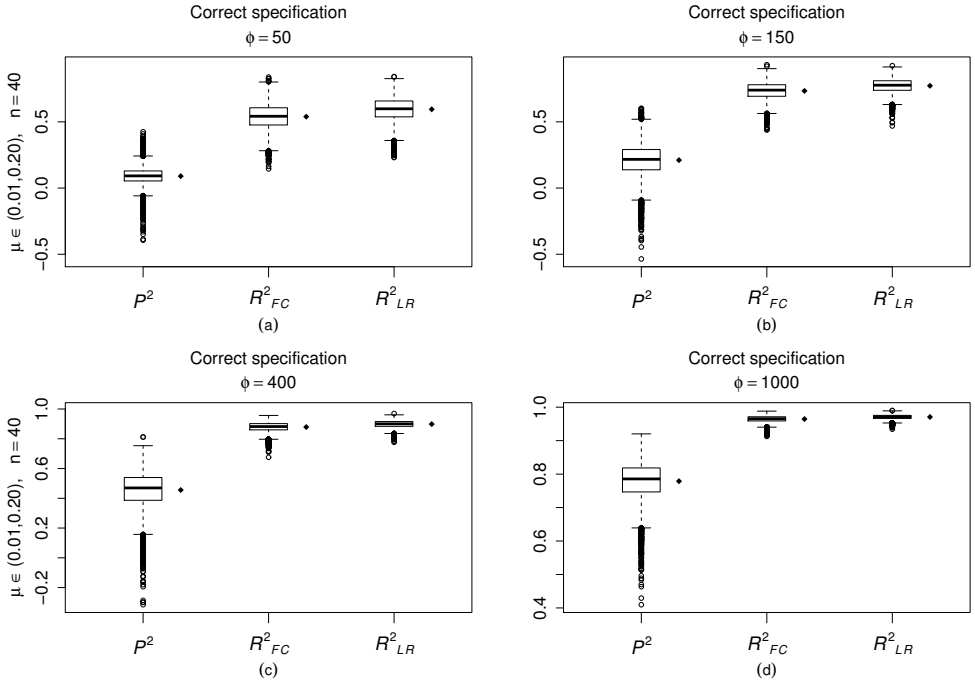

Figure 3. Correct specification: $g(\mu_t) = \beta_1 + \beta_2\, x_{t2} + \beta_3\, x_{t3} + \beta_4\, x_{t4} + \beta_5\, x_{t5}$; sample size $n = 40$. (**a**) $\phi = 50$; (**b**) $\phi = 150$; (**c**) $\phi = 400$; (**d**) $\phi = 1000$.

To summarize, at the outset, we shall consider the response mean around 0.5. When the model is correctly specified, the P^2 have their values close to one, especially when the model precision or sample size increases. When the proposed model omits important covariates, the P^2 values tend to depart considerably of one and stay below 0.5. The measure R^2_{FC} and R^2_{LR} present similar behavior. On the other hand, when the mean of the response is concentrated near zero or one, the P^2 values differ considerably from one, taking negative values even when the model is correctly specified, revealing as it is difficult make prediction close to the boundaries of the unit interval.

Indeed, scenarios in which the model present large dispersion and a substantial concentration of values on one of the boundaries of the standard unit interval tend to present influential observations. In these situations, Ref. [22] argue that for the beta regression models the ML parameter estimation based on the BFGS nonlinear optimization algorithm proved to be typically not robust in influential cases. The P^2 criterion is based on the PRESS-like statistic which presents a relationship both with residuals and influence measures. In this sense, this new criterion that we proposed for the beta regression models outperforms the R^2_{LR} and R^2_{FC} in identifying problems on fit the model when $\mu \approx 0$ or $\mu \approx 1$ and the precision is not so large. However, this fact does not disable the use of R^2-like statistics. The P^2 criterion can be viewed as a measure of model bias whereas the R^2 is a quantifier of the model variance. What we emphasize is that we must also consider the P^2 criterion to select the model that best fit a dataset. In the applications we shall present results that show as the R^2 and P^2 criteria contain different and important information about the model-fitting.

Another important question is the link function to the mean sub-model. All simulations, we carried out until now were based on logit link function. In what follows, we present Monte Carlo simulation results in which we consider other link functions, namely: probit, complementary log–log, log–log, and Cauchy, respectively defined as $g(\mu) = \Phi^{-1}(\mu)$, $g(\mu) = -\log\{-\log(\mu)\}$, $g(\mu) = \log\{-\log(1-\mu)\}$ and $g(\mu) = \tan\{\pi(\mu - 0.5)\}$. It is important to emphasize that the same link function is used both to generate the response observations and to fit the model. Our goal is to evaluate the performance of each link function on different ranges of mean and dispersion response, in special we aim to identifying if the link function is related to the problems in fitting the beta regression model when the response is close of the boundaries of the standard unit interval. Thus, we must fit the model correctly.

The results presented in Table 2 showed that when the response mean is close to one, the use of complementary log–log function leads to models with better predictive power as well as better goodness-of-fit. On the other hand, if the mean is close to zero the best results are provided by the log–log link function. When the mean is scattered on the standard unit interval both the probit and logit functions perform well. The Cauchy model performance well only when $\mu \in (0.20, 0.80)$. Thereby, we can deduce that the link function is related with the small values the P^2 criteria when $\mu \approx 0$ and $\mu \approx 1$ displayed in Table 1, since all scenarios were fitted by the logit model. Thus, the appropriate link function can improve the robustness of the ML estimation procedure of the beta models in the presence of influential points. It is noteworthy that these conclusions are supported on the P^2 criterion.

Table 2. Mean values of the statistics. True model: $g(\mu_t) = \beta_1 + \beta_2\, x_{t2} + \beta_3\, x_{t3} + \beta_4\, x_{t4} + \beta_5\, x_{t5}$, $x_{ti} \sim U(0,1)$, $i = 2,3,4,5$, $t = 1,\ldots,n$ and ϕ constant across observations.

	$\mu \to$	$\mu \in (0.005, 0.12)$			$\mu \in (0.20, 0.88)$			$\mu \in (0.90, 0.99)$		
n	$\phi \to$	20	150	400	20	150	400	20	150	400
		Probit								
40	P^2	0.281	0.561	0.759	0.610	0.913	0.966	0.282	0.562	0.759
	P_c^2	0.222	0.525	0.739	0.578	0.905	0.963	0.222	0.525	0.739
	R^2_{LR}	0.392	0.792	0.910	0.599	0.912	0.966	0.394	0.792	0.910
	$R^2_{LR_c}$	0.341	0.774	0.903	0.566	0.905	0.963	0.344	0.775	0.903
	R^2_{FC}	0.389	0.791	0.910	0.599	0.912	0.965	0.391	0.791	0.910
	$R^2_{FC_c}$	0.338	0.773	0.902	0.565	0.905	0.963	0.340	0.773	0.902
		C-Log-Log								
40	P^2	0.109	0.195	0.343	0.535	0.883	0.953	0.370	0.694	0.851
	P_c^2	0.035	0.128	0.288	0.497	0.873	0.949	0.317	0.668	0.839
	R^2_{LR}	0.370	0.780	0.905	0.574	0.903	0.962	0.333	0.741	0.884
	$R^2_{LR_c}$	0.260	0.741	0.888	0.499	0.886	0.955	0.216	0.696	0.863
	R^2_{FC}	0.362	0.773	0.902	0.574	0.903	0.962	0.327	0.739	0.883
	$R^2_{FC_c}$	0.308	0.754	0.893	0.539	0.895	0.959	0.271	0.717	0.873
		Log-Log								
40	P^2	0.370	0.694	0.851	0.536	0.883	0.953	0.109	0.196	0.342
	P_c^2	0.318	0.668	0.839	0.497	0.873	0.949	0.034	0.129	0.287
	R^2_{LR}	0.334	0.742	0.884	0.574	0.903	0.962	0.370	0.780	0.905
	$R^2_{LR_c}$	0.279	0.720	0.874	0.538	0.895	0.959	0.318	0.762	0.897
	R^2_{FC}	0.327	0.739	0.883	0.574	0.904	0.962	0.362	0.774	0.902
	$R^2_{FC_c}$	0.271	0.718	0.873	0.539	0.896	0.959	0.308	0.755	0.893
		Cauchy								
40	P^2	0.031	0.158	0.330	0.511	0.877	0.951	0.031	0.158	0.330
	P_c^2	0.006	0.136	0.313	0.498	0.874	0.949	0.006	0.136	0.313
	R^2_{LR}	0.183	0.556	0.769	0.599	0.913	0.966	0.182	0.557	0.768
	$R^2_{LR_c}$	0.161	0.545	0.763	0.588	0.911	0.965	0.161	0.546	0.762
	R^2_{FC}	0.080	0.460	0.702	0.561	0.905	0.963	0.081	0.460	0.702
	$R^2_{FC_c}$	0.057	0.446	0.694	0.549	0.903	0.962	0.057	0.446	0.694

3.2. Linear Setting: Varying Dispersion

In this section, we shall report simulation results to beta regression models with varying dispersion. All results were obtained using 10,000 Monte Carlo replications. Under model misspecification, the true data generating process considers varying dispersion, but a fixed dispersion beta regression is estimated. We also used different covariates in the mean and precision sub-models. The sample sizes are $n = 40, 80, 120$. We generated 40 values for each covariate and replicated them, once, twice, and three times, respectively, to get covariate values for $n = 80$ and $n = 120$. Using this procedure, the intensity degree of nonconstant dispersion $\lambda = \max\{\phi_1,\ldots,\phi_n\}/\min\{\phi_1,\ldots,\phi_n\}$ remains constant as the sample size changes. The numerical results were obtained using the following beta regression model: $g(\mu_t) = \log(\mu_t/(1-\mu_t)) = \beta_1 + \beta_i\, x_{ti}$, and $\log(\phi_t) = \gamma_1 + \gamma_i\, z_{ti}$, $x_{ti} \sim U(0,1)$, $z_{ti} \sim U(-0.5, 0.5)$, $i = 2,3,4,5$, and $t = 1,\ldots,n$ under different choices of parameters (Scenarios): Scenario 5: $\beta = (-1.3, 3.2)^\top$,

$\mu \in (0.22, 0.87)$, $[\gamma = (3.5, 3.0)^\top; \lambda \approx 20]$, $[\gamma = (3.5, 4.0)^\top; \lambda \approx 50]$ and $[\gamma = (3.5, 5.0)^\top; \lambda \approx 150]$. Scenario 6: $\beta = (-1.9, 1.2, 1.6, 2.0)^\top$, $\mu \in (0.24, 0.88)$, $[\gamma = (2.4, 1.2, -1.7, 1.0)^\top; \lambda \approx 20]$, $[\gamma = (2.9, 2.0, -1.7, 2.0)^\top; \lambda \approx 50]$ and $[\gamma = (2.9, 2.0, -1.7, 2.8)^\top; \lambda \approx 150]$. Finally, Scenarios 7 and 8 (Full models): $\beta = (-1.9, 1.2, 1.0, 1.1, 1.3)^\top$, $\mu \in (0.20, 0.88)$, $[\gamma = (3.2, 2.5, -1.1, 1.9, 2.2)^\top; \lambda \approx 20]$, $[\gamma = (3.2, 2.5, -1.1, 1.9, 3.2)^\top; \lambda \approx 50]$, and $[\gamma = (3.2, 2.5, 1.1, 1.9, 4.0)^\top; \lambda \approx 200]$. Please note that Scenarios 7 and 8 present the same generation data process. However, in Scenario 7 the dispersion is estimated as a constant (misspecification) and in Scenario 8 the dispersion is correctly modeled.

In Table 3, we present the mean values for 10,000 statistic replications. In this table, we report the results only for $n = 40$. Next, we presented boxplots for the 10,000 statistic replications to other sample sizes. We are considering μ close to 0.5. We notice based on Table 3 that under model misspecification the statistics display smaller values in comparison with Scenario 8 (correct specification), in which as greater is λ greater are the values of the statistics, as expected. When the dispersion is postulated as fixed, as the intensity degree of nonconstant dispersion increases, the mean values of the statistics decreases, which correctly points out for the model misspecification. It is noteworthy that under correct model specification the values of three statistics are so different. In special the P^2 values are greater than the values of R^2-like criteria. Furthermore, the values of the R^2_{FC} are considerably smaller than the values of the R^2_{LR}, in special when λ increases. For example, taking $\lambda = 20, 50, 200$, $n = 40$ we have $R^2_{LR} = (0.796, 0.816, 0.840)$ and $R^2_{FC} = (0.649, 0.627, 0.500)$ (Table 3–Scenario 8). Figure 4 supports this evidence. When λ and the sample size increase, for example $n = 80$ and $n = 120$, the values of P^2 criterion tend to concentrate close to one, whereas the values of R^2_{LR} and R^2_{FC} tend to concentrate below 0.8 and 0.6, respectively.

Table 3. Mean values of the statistics. Misspecified models, ϕ fixed: Scenarios 5, 6 and 7 versus Scenario 8 (correct specification), n = 40.

	Scenario 5			**Scenario 6**			**Scenario 7**			**Scenario 8**		
True models	$g(\mu_t) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3}$ $h(\phi_t) = \gamma_1 + \gamma_2 z_{t2} + \gamma_3 z_{t3}$			$g(\mu_t) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4}$ $h(\phi_t) = \gamma_1 + \gamma_2 z_{t2} + \gamma_3 z_{t3} + \gamma_4 z_{t4}$			$g(\mu_t) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 x_{t5}$ $h(\phi_t) = \gamma_1 + \gamma_2 z_{t2} + \gamma_3 z_{t3} + \gamma_4 z_{t4} + \gamma_5 z_{t5}$			$g(\mu_t) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 x_{t5}$ $h(\phi_t) = \gamma_1 + \gamma_2 z_{t2} + \gamma_3 z_{t3} + \gamma_4 z_{t4} + \gamma_5 z_{t5}$		
Estimated models	$g(\mu_t) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3}$			$g(\mu_t) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4}$			$g(\mu_t) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 x_{t5}$			$g(\mu_t) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 x_{t5}$ $h(\phi_t) = \gamma_1 + \gamma_2 z_{t2} + \gamma_3 z_{t3} + \gamma_4 z_{t4} + \gamma_5 z_{t5}$		
$\lambda \rightarrow$	20	50	150	20	50	150	20	150	150	20	50	200
P^2	0.759	0.718	0.674	0.545	0.565	0.523	0.638	0.624	0.529	0.885	0.906	0.914
P^2_c	0.739	0.695	0.647	0.493	0.515	0.469	0.585	0.569	0.460	0.851	0.878	0.888
R^2_{LR}	0.782	0.743	0.700	0.580	0.611	0.577	0.670	0.653	0.554	0.796	0.816	0.840
$R^2_{LR_c}$	0.764	0.722	0.675	0.532	0.567	0.529	0.622	0.602	0.488	0.735	0.761	0.792
R^2_{FC}	0.777	0.735	0.688	0.553	0.588	0.548	0.668	0.648	0.531	0.649	0.627	0.500
$R^2_{FC_c}$	0.759	0.713	0.662	0.502	0.541	0.497	0.620	0.597	0.462	0.544	0.515	0.350

We shall focus on $n = 40$, Figure 4e the true intensity degree of nonconstant dispersion is close to 200, with $\phi_{\max} \approx 260$ and $\phi_{\min} \approx 2$, whereas $\widehat{\lambda} \approx 400$ with $\widehat{\phi}_{\max} = 730$ and $\widehat{\phi}_{\min} \approx 1.9$, that is a substantial distortion of the true intensity degree of nonconstant dispersion. Indeed, it is a substantial distortion of the true variance of the response observations.

Since the R^2-like criteria, select the model that can better explain the variability of the response, it is plausible that these measures present lower values when the distortions between the true and estimated variances of the response variable are so large. Please note that the $R^2_{LR_c}$ takes several values smaller than 0.6 and the $R^2_{FC_c}$ even takes negative values whereas overall the values of P^2_c are greater than 0.6 (Figure 4e). Additionally, in this sense the R^2_{FC} criterion proved to be more rigorous than the R^2_{LR} criterion. This is a strong evidence that models with small R^2_{FC} and high R^2_{LR} values are worthy of further investigation. Indeed, the best fitted model should display high and close values of the three criteria and of their penalized versions.

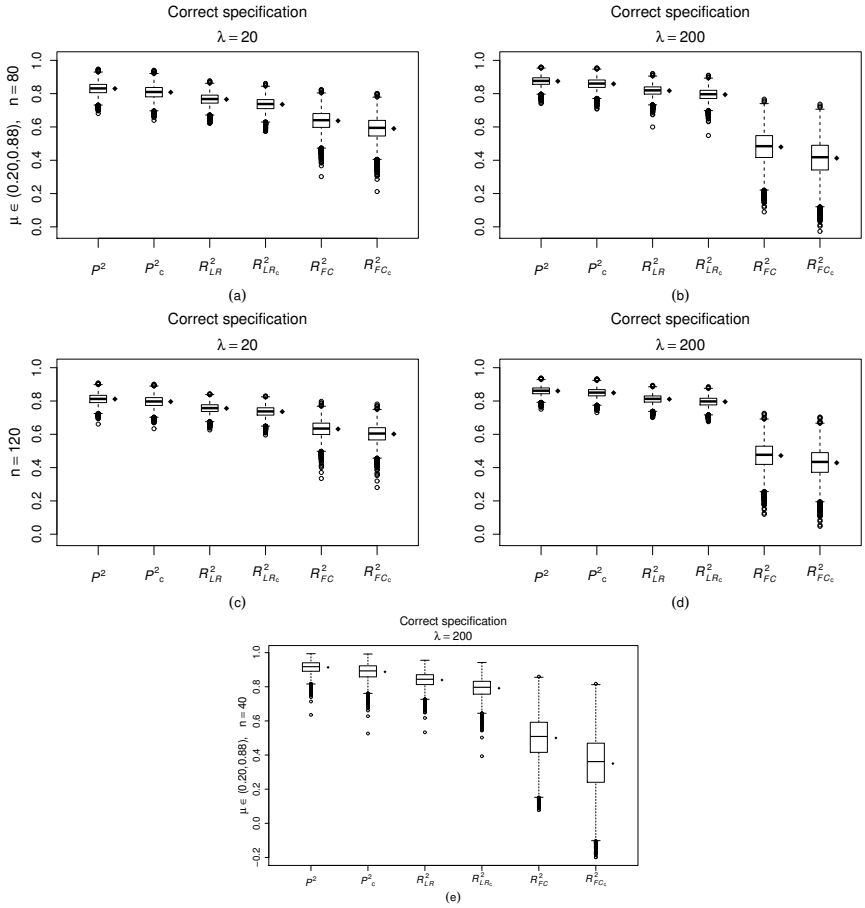

Figure 4. Correct specification: $g(\mu_t) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 x_{t5}$, $h(\phi_t) = \gamma_1 + \gamma_2 z_{t2} + \gamma_2 z_{t2} + \gamma_3 z_{t3} + \gamma_4 z_{t4} + \gamma_5 z_{5t}$. **(a)** $\mu \in (0.20, 0.80), n = 80, \lambda = 20$; **(b)** $\mu \in (0.20, 0.80), n = 80, \lambda = 200$; **(c)** $\mu \in (0.20, 0.80), n = 120, \lambda = 20$; **(d)** $\mu \in (0.20, 0.80), n = 120, \lambda = 200$; **(e)** $\mu \in (0.20, 0.80), n = 40, \lambda = 200$.

3.3. Nonlinear Setting

In what follows, we shall present Monte Carlo experiments for the class of nonlinear beta regression models. The numerical results were obtained using the following beta regression model as data generating processes:

$$\log\left(\frac{\mu_t}{1-\mu_t}\right) = \beta_1 + x_{t2}^{\beta_2} + \beta_3 \log(x_{t3} - \beta_4) + \frac{x_{t3}}{\beta_5}, \quad t = 1, \ldots, n,$$

$x_{t2} \sim U(1,2)$, $x_{t3} \sim U(4.5, 34.5)$ and ϕ were kept fixed throughout the experiment. Here we use the starting value scheme for the estimation by ML proposed by [23]. The precision and the sample size are respectively $\phi = (20, 50, 150, 400)$, $n = (20, 40, 60)$. Additionally, $\beta = (1.0, 1.9, -2.0, 3.4, 7.2)^\top$ that yields $\mu \in (0.36, 0.98)$. To evaluate the criterion performance on account of nonlinearity negligence, we consider the following model specification: $\log\left(\frac{\mu_t}{1-\mu_t}\right) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3}$. All results are based on 10,000 Monte Carlo replications and for each replication.

We evaluated the behavior of the statistics both under model misspecification and under model correct specification. The results displayed in Table 4 reveal that all statistics present values smaller when the model is missspecified. For example, fixing the precision value of $\phi = 400$, for $n = 20$, we have values of P^2, R^2_{LR} and R^2_{FC} equal to $0.576, 0.700, 0.637$, respectively. For $n = 40$ and $n = 60$ the values of the statistics are $0.568, 0.698, 0.634$ and $0.562, 0.698, 0.633$, respectively. We simulated other nonlinear patterns to the sub-model mean predictor, and in some simulations the three criteria did not present smaller values of the feasible linear model than to the nonlinear model correctly specified.

Table 4. Mean values of the statistics. True model: $g(\mu_t) = \beta_1 + x_{t2}^{\beta_2} + \beta_3 \log(x_{t3} - \beta_4) + \frac{x_{t3}}{\beta_5}$, $x_{t2} \sim U(1,2)$, $x_{t3} \sim U(4.5, 34.5)$, $\beta = (1.0, 1.9, -2.0, 3.4, 7.2)^\top$, $\mu \in (0.36, 0.98)$, $t = 1, \ldots, n$, ϕ fixed. Misspecification: $g(\mu_t) = \beta_1 + \beta_2\, x_{t2} + \beta_3\, x_{t3}$ (omitted nonlinearity).

Estimated Model	With misspecification: $g(\mu_t) = \beta_1 + \beta_2\, x_{t2} + \beta_3\, x_{t3}$												Correctly		
n	20				40				60				60		
$\phi \rightarrow$	20	50	150	400	20	50	150	400	20	50	150	400	50	150	400
P^2	0.485	0.535	0.564	0.576	0.438	0.508	0.550	0.568	0.420	0.496	0.543	0.562	0.849	0.936	0.975
P^2_c	0.388	0.448	0.483	0.497	0.391	0.467	0.513	0.532	0.388	0.469	0.518	0.539	0.835	0.930	0.973
R^2_{LR}	0.578	0.647	0.684	0.700	0.563	0.639	0.681	0.698	0.557	0.636	0.680	0.698	0.883	0.953	0.982
$R^2_{LR_c}$	0.499	0.581	0.625	0.643	0.526	0.608	0.654	0.673	0.533	0.616	0.662	0.681	0.863	0.945	0.979
R^2_{FC}	0.486	0.574	0.619	0.637	0.448	0.556	0.612	0.634	0.437	0.550	0.609	0.633	0.879	0.951	0.981
$R^2_{RC_c}$	0.389	0.494	0.548	0.569	0.402	0.519	0.580	0.604	0.407	0.526	0.588	0.613	0.867	0.946	0.979

4. Applications

4.1. Fluid Catalytic Cracking

The first application employs real data from the graduation work of [24], from Chemistry Department of the Colombia National University. It is based on the Fluid Catalytic Cracking (FFC) process, considered the heart of a gasoline refinery. [24] explains that the FCC process is used to convert hydrocarbons of high molecular weight into small molecules of high commercial value, through the contact of hydrocarbons with a catalyst. The zeolite USY is the major catalyst of the process. The FCC process also involves the vanadium element, steam, and temperature. However, the vanadium on the catalyst decreases gasoline production.

Is special, the vanadium affects the crystallinity of zeolite USY depending on steam concentration and of the temperature during the process. The aim here is modeling the percentage of crystallinity of zeolite USY (y), based on different concentrations of vanadium (x_2) and steam (x_3), and two values of the process temperature (x_4). Typically, the higher the vanadium and steam concentrations, the lower the percentage of crystallinity. [22] modeled these data. At the outset, the authors fitted several linear beta regression models and carried out the residual analysis which made clear the nonlinear trend. Thus, the authors modeling these data using a logit nonlinear beta model defined as

$$g(\mu_t) = \beta_1 + \beta_2 x_{t2}/(x_{t2} + \beta_3) + \beta_3 x_{t3} + \beta_4 \sqrt{x_{t4}} \quad \text{and} \quad \log(\phi_t) = \gamma_1 + \gamma_2 x_{t4}^2, \tag{20}$$

$t = 1 \ldots, 28$. We fitted the model in (20) considering five link functions, namely: logit, probit, log–log, complementary log–log (C-Log-Log) and Cauchy. We shall present only the logit and complementary log–log model inferences (Table 5). Similar results are obtained by use the probit, log–log, and Cauchy link functions. However, we report that the parameter γ_2 was significantly different from zero, at the usual nominal levels, only for the model with the Cauchy link function. On the other hand, to the models with C-Log-Log and logit link functions the parameter γ_2, is far to be significantly different from zero, p-value equal to 0.404 and 0.200, respectively.

Table 5. Parameter estimates, standard errors (s.e.), relative changes in estimates and in standard errors due to cases exclusions and respective p-values. Varying dispersion model.

Model	**C–Log–Log**							**Logit**						
	β_1	β_2	β_3	β_4	β_5	γ_1	γ_2	β_1	β_2	β_3	β_4	β_5	γ_1	γ_2
Full	0.90	−0.05	−26.9	−0.16	−0.32	4.05	0.26	2.29	−0.10	−26.9	−0.29	−0.68	3.92	0.40
data	0.07	0.02	4.34	0.06	0.06	0.37	0.31	0.15	0.04	4.36	0.11	0.13	0.37	0.31
	0.000	0.006	0.000	0.003	0.000	0.000	**0.404**	0.000	0.010	0.000	0.007	0.000	0.000	**0.200**
	Estimate and standard error changes (%) and p-values.													
Obs.	2.1	−16.8	5.9	−30.4	2.6	9.5	−46.7	2.7	−16.6	6.6	−34.6	5.4	9.4	−25.9
10,20	−13.1	10.8	0.2	−3.4	−7.6	7.8	13.5	−7.4	13.8	0.5	1.1	−4.6	7.8	13.5
22,28 del.	0.000	0.040	0.000	0.033	0.000	0.000	0.695	0.000	0.062	0.000	0.086	0.000	0.000	0.404
Obs.	−0.7	−1.2	2.5	−13.0	−2.8	−0.6	−31.5	1.7	−7.1	11.2	−54.6	9.1	−8.7	121.8
13,20	3.7	16.5	8.1	15.7	10.0	1.0	12.5	5.3	26.6	−23.9	−4.3	10.3	1.0	12.7
23,27 del.	0.000	0.020	0.000	0.025	0.000	0.000	0.612	0.000	**0.061**	0.000	**0.210**	0.000	0.000	0.012

Nevertheless, we computed the selection criteria for the five models and presented in Table 6. The results in this table evidence that the values of the statistics are overall low, except to complementary log–log model. Furthermore, the lower values of the P^2 statistic when compared with the values of R^2-like criteria, is special to logit model, is an indication of same misspecification on the fitted models. Thus, from now on we shall focus on the complementary log–log and logit models.

In what follows, we shall perform residual and influence diagnostics for the fitted models based on (20) and using the logit and complementary log–log link functions, see Figures 5 and 6, respectively. The index plots of the Cook-like distances identify the observations $\{10, 12, 16, 24\}$ as influential, for the two link function models. Furthermore, the case $\{20\}$ is worthy of further investigation for logit model, Figure 5c,d. However, the most important information is provided by the normal probability plot for the logit model in Figure 5b. Here there are two points on the boundaries of envelope bands, cases 22 and 28. Typically, these are influential cases.

Table 6. Criterion values. Nonlinear Model. Data on FCC.

Criteria	Constant Dispersion					Varying Dispersion				
	Logit	Probit	Log–Log	C–Log–Log	Cauchy	Logit	Probit	Log–Log	C–Log–Log	Cauchy
P^2	0.42	0.55	0.22	0.70	0.09	0.20	0.33	0.31	0.66	0.51
P_c^2	0.29	0.44	0.05	0.63	−0.1	−0.03	0.14	0.12	0.59	0.37
R^2_{FC}	0.68	0.68	0.67	0.68	0.51	0.67	0.68	0.66	0.69	0.49
$R^2_{FC_c}$	0.60	0.61	0.59	0.61	0.40	0.57	0.59	0.56	0.59	0.34
R^2_{LR}	0.69	0.69	0.68	0.70	0.65	0.70	0.71	0.70	0.71	0.70
$R^2_{LR_c}$	0.55	0.55	0.54	0.56	0.49	0.55	0.59	0.55	0.56	0.55

Based on the above analysis, we removed from the data combinations of the cases 10, 12, 16, 20, 22, 24, 28 (residual/Cook plots) and 13, 18, 23, 27 singled out additionally by the local influence plots (Figure 6)). In Figure 6 we carried out the local influence analysis based on the perturbation simultaneous of the covariate vanadium, which is present both in the mean and dispersion predictor.

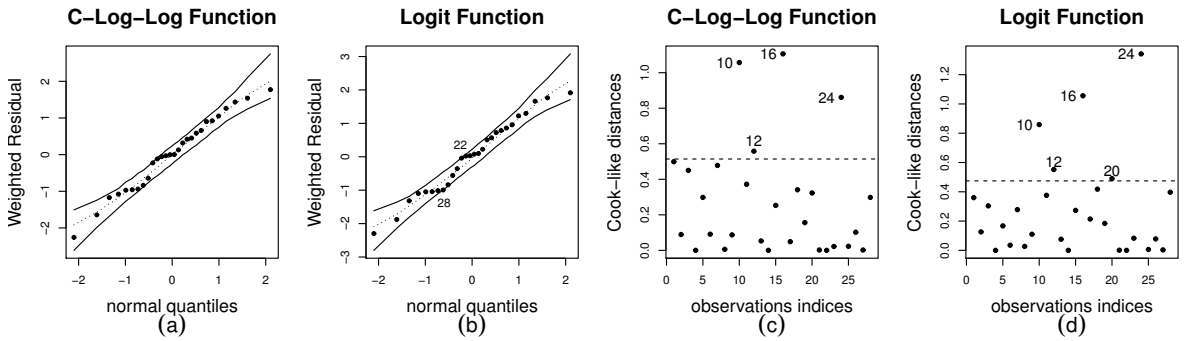

Figure 5. Residual and Cook-like distance plots. Varying dispersion model. Data on FCC. (**a**) Envelope band of weight residual and link function C-Log-Log; (**b**) Envelope band of weight residual and link function Logit; (**c**) Cook-like distance and link function C-Log-Log ; (**d**) Cook-like distance and link function C-Log-Log.

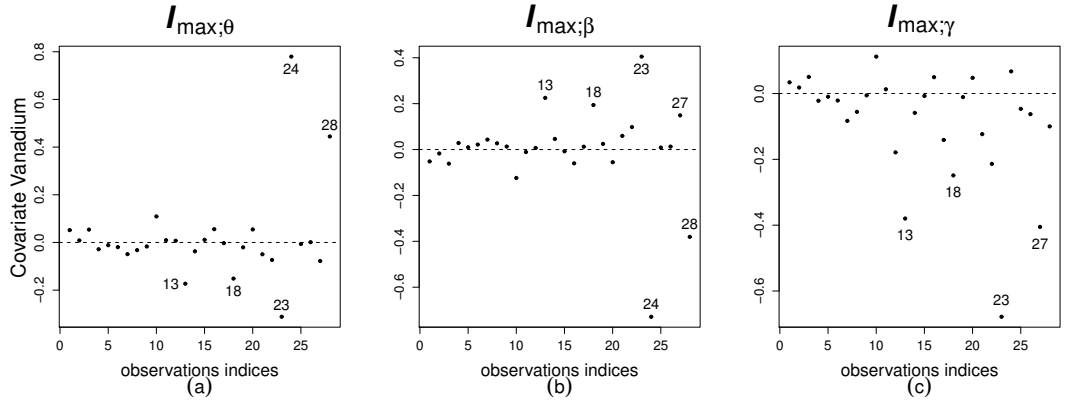

Figure 6. Local influence plots. Varying dispersion model. Data on FCC. (**a**) Simultaneous perturbation; (**b**) Mean perturbation; (**c**) Dispersion perturbation.

Thus, we fit the models after the exclusions. We take advantage of the information in Table 5, where we present the relative changes (%) in parameter estimates and standard error estimates, as well as the p-values after the exclusions that most affected the model-fitting. From Table 5 we note that the estimation process of the complementary log–log model proved to be more robust to influential cases than the estimation process of the logit model.

The set $\{10, 20, 22, 28\}$ impacts the estimates of β_2 and β_4 for both models. However, the complementary log–log model ensures the same inference conclusions whereas to the logit model these parameters become non-significantly different from zero at the 5% level, in special β_4(p-value = 0.086). For the logit model, the set $\{13, 20, 23, 27\}$ is still more influential. The exclusion of this set strongly affects the estimates of γ_2, such that the dispersion becomes varying and β_4 and β_2 become non-significant at the 20% and 5% levels, respectively. This fact is a strong evidence that the parameter estimates of the full data logit model are biased. The P^2 selection criterion was able in identifying this bias pattern, what explains the low value of this criterion to the logit model, even reaching a negative value for their penalized version. Furthermore, this bias pattern is due to the non-robustness of the ML estimation process in the presence of influential cases. We note also that the cases 20 and 28 highlighted on the limits of envelope bands proved being influential cases revealing the importance of to evaluate this plot carefully. Thus, the nonlinear model with varying dispersion does not seem to be a good option to these data. On the other hand, the fit of a nonlinear model with fixed precision, based on complementary log–log link function presents satisfactory values of the all selection criteria (Table 6).

The residual plots in Figure 7a,b support this conclusion, since to the complementary log–log model all residuals are randomly scattered within the envelope bands whereas to the logit model there is the case 5 as potentially influential. However, the cases highlighted as worthy of further investigation by the total local influence did not change the inference conclusions, is despite to yield greater changes in the parameter estimates of the logit model than to the complementary log–log model. Figure 7c,d. For this data set, $y \in (0.64.0.96)$ with a median close to 0.81, and the estimated values of ϕ are quite similar for all link function models, close to 65. In this scenario we verified by Monte Carlo simulation that the models based on complementary log–log functions provide the highest values of all selection criteria. Thus, the application only confirms the simulation results.

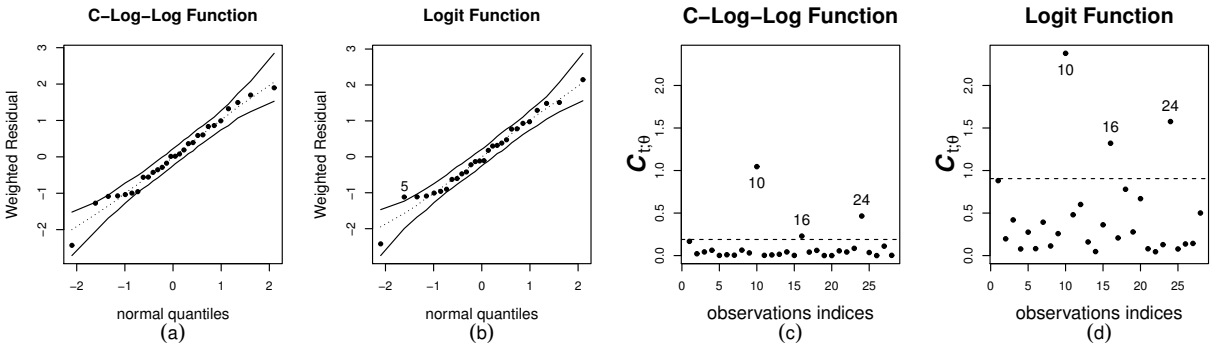

Figure 7. Residual and total local influence plots. Constant dispersion. Data on FCC. (**a**) Envelope band of weight residual and link function C-Log-Log; (**b**) Envelope band of weight residual and link function Logit; (**c**) Cook-like distance and link function C-Log-Log ; (**d**) Cook-like distance and link function C-Log-Log.

4.2. Simultaneity Factor

The second application relates to the distribution of natural gas for home usage in São Paulo, Brazil. The real data were obtained from the Instituto de Pesquisas Tecnológicas-IPT (https://www.ipt.br/) and

the Companhia de Gás de São Paulo-COMGÁS (https://www.comgas.com.br/). The response variable (y) is the simultaneity factor, the covariate x_2 is the log of computed power and the sample size is $n = 42$. Ref. [25] built a bootstrap-based prediction interval for the response variable-based on beta regression model with constant dispersion defined as $\log(\mu_t/(1-\mu_t)) = \beta_1 + \beta_2\, x_{t2}$, which was selected by the classical version of PRESS statistic given by $\text{PRESS}_C = \sum_{t=1}^{42}(y_t - \widehat{y}_{(t)})^2/42$. Here we aim at selecting the best model to the data on simultaneity factor using the P^2, R^2_{LR} and R^2_{FC} criteria. We consider five link functions for μ sub-model, namely: logit, probit, log–log, complementary log–log and Cauchy. Thus, we fitted five beta regression models based on

$$g(\mu_t) = \beta_1 + \beta_2\, x_{t2} \quad \text{and} \quad h(\phi_t) = \beta_1 + \beta_2\, x_{t2}, \quad t = 1, \ldots, 42. \tag{21}$$

We also fitted beta regression models with constant dispersion based on the same five link functions. For the logit model the maximum likelihood parameter estimates are $\widehat{\beta}_1 = -1.71$, $\widehat{\beta}_2 = -0.33$ and $\widehat{\phi} \approx 79$ ($\widehat{\beta}_1 = -0.61$, $\widehat{\beta}_2 = -0.33$, and $\widehat{\phi} \approx 71$, log–log model). The Figure 8a shows that $y \in [0.016, 0.464]$, and the median is equal to 0.069.

We highlight that the simulation results obtained under a similar scenario favor the log–log models. These achievements are provided by the values displayed in the Table 7 (Constant dispersion). The values in this table also reveal that using varying dispersion models improve the fit. Here, the evidence that to fit this data the log–log link function is the best choose and a varying dispersion model is the best model are support only by the P^2 criterion. We shall carry out an influential and residual analysis to prove the outperformance of the P^2 criterion in this situation.

Table 7. Criteria values. Data on simultaneity factor.

	Constant Dispersion					**Varying Dispersion**				
Criteria	Logit	Probit	Log–Log	C–Log–Log	Cauchy	Logit	Probit	Log–Log	C–Log–Log	Cauchy
P^2	0.42	0.50	0.65	0.28	−1.67	0.70	0.83	0.88	0.62	−0.98
P^2_c	0.39	0.47	0.64	0.24	−1.81	0.67	0.81	0.87	0.59	−1.13
R^2_{FC}	0.69	0.71	0.72	0.68	0.39	0.69	0.71	0.72	0.68	0.39
$R^2_{FC_c}$	0.67	0.70	0.71	0.66	0.36	0.67	0.69	0.70	0.65	0.35
R^2_{LR}	0.72	0.71	0.69	0.73	0.67	0.74	0.74	0.74	0.74	0.67
$R^2_{LR_c}$	0.69	0.68	0.65	0.69	0.62	0.71	0.71	0.71	0.71	0.62

In the Figure 8b,c we present the Cook-like distance index plots for the constant and varying dispersion models based on log–log link function, respectively. These plots shown that to the beta regression model with constant dispersion the coefficient estimates of the mean sub-model are highly sensitive to the case 21, whereas the potential influence of this case is set aside when the dispersion is modeled. This is a strong evidence that the varying dispersion model fits better the data. Forward, we shall focus on the beta regression with varying dispersion, in the Table 8 we present the inference results for the models fitted considering the five link functions. For the logit and complementary log–log functions, z_2 is only significative at a 10% level, whereas when we are using the Cauchy link function this covariate is no longer significant (p-value = 0.5133). When we use the probit function, the covariate becomes significant at the level of 5%. However, the most significance level for z_2 is only reached when the fit considers the log–log function (p-value = 0.0088).

Table 8. Parameter estimates, standard errors (s.e.) and p-values. Data on simultaneity factor.

Par.	Logit			Probit			Log–Log			C–Log–Log			Cauchy		
	Estim.	**s.e.**	**p-val.**	**Estim.**	**s.e.**	**p-val.**	**Estim.**	**s.e.**	**p-val.**	**Estim.**	**s.e.**	**p-val.**	**Estim.**	**s.e.**	**p-val.**
β_1	-1.72	0.09	0.000	-1.01	0.05	0.000	-0.63	0.05	0.000	-1.82	0.08	0.000	-2.51	0.18	0.000
β_2	-0.80	0.08	0.000	-0.41	0.05	0.000	-0.31	0.04	0.000	-0.74	0.07	0.000	-1.47	0.14	0.000
γ_1	4.00	0.33	0.000	3.91	0.33	0.000	3.81	0.33	0.000	4.05	0.33	0.000	4.01	0.30	0.000
γ_2	0.54	0.30	0.067	0.65	0.30	0.027	0.77	0.29	0.009	0.50	0.30	0.096	0.19	0.29	0.513

In the Figure 8d,e we present the normal probability plots with simulated envelopes to varying dispersion models based on the log–log, the logit, and the Cauchy link functions, respectively. These plots reveal that the log–log model yields the best fit whereas the Cauchy model yields the worst fit. This is the same conclusion provides by the P^2 criterion, whereas by the R^2_{LR} criterion all link functions could provide a good fit, even the Cauchy link function. For instance, to the logit and Cauchy models the $(R^2_{LR}, R^2_{LR_c})$ are, respectively, $(0.74, 0.71)$ and $(0.67, 0.62)$.

The performance of the R^2_{LR} is proved to be poor when we look the Figure 8f, which clarify unmistakably lack of fit of the Cauchy regression. Even the selection of the logit model would not be appropriate since there are ranges of residuals not randomly distributed across the envelope bands (Figure 8e). We note that the P^2 reaches negative values and the R^2_{FC} is able in identifying some problem on the model variability, whether the Cauchy function is used ($R^2_{FC} = 0.39$). Whether a practitioner does not take into account the other statistics beyond of the R^2_{LR} criterion one could select both logit and Cauchy model to fit the data. This conclusion would be quite counter for what is proved by residual plots and inference results. Please note that the R^2_{LR} criterion presents a close relation with AIC-like criteria. Thus, we must be careful in using a criterion to choose a model even the usual and classical criteria.

Although we must emphasize that P^2-like criteria must be used jointly with the R^2-like criteria. We shall focus on the normal probability plot of the log–log fit (Figure 8d). It is possible to note two points out of the envelope bands just as a slight linear tendency on the residual distribution close of these two points. This pattern explains the discrepancy between the values of the R^2-like criteria and the value of P^2 criterion, which are equal 0.7, and 0.9, respectively (Table 7). This pattern suggests some problem in the dispersion sub-model or in the distribution of probability postulated for the response. Thus, we decide fit other beta regression models, considering different link functions also for the dispersion sub-model, just as different functions for the computed power beyond of the logarithm function, as covariates. The best fit is still the one provides by the beta regression model defined in (21) considering the log–log link function. In the future we can consider other distributions to fit this data as the simplex distribution that is a dispersion model and can provide a better fit.

However, the beta regression model defined in (21) and based on log–log function is useful for modeling the data on simultaneity factor. This example clarifies how it is important to consider both prediction criteria and different versions of the R^2 criteria to select the best model to fit a dataset.

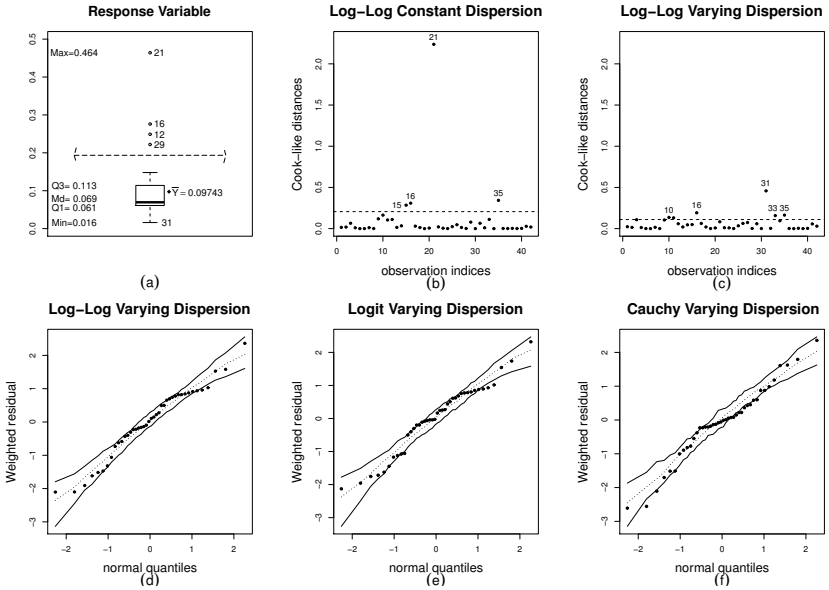

Figure 8. Diagnostic plots. Data on simultaneity factor. (**a**) Boxplot of the response; (**b**) Cook-like distance and link function Log-Log with constant dispersion; (**c**) Cook-like distance and link function Log-Log with varying dispersion; (**d**) Envelope band of weight residual and link function Log-Log with varying dispersion; (**e**) Envelope band of weight residual and link function Logit with varying dispersion; (**f**) Cook-like distance and link function Cauchy with varying dispersion.

5. Conclusions and Future Work

In this paper, we develop the P^2 criterion based on the PRESS-like machine learning tool for the class of beta regression models. We presented results of Monte Carlo simulations carried out to evaluate the performance the P^2 criterion and of the versions R^2_{LR} and R^2_{FC} of the R^2 criterion, under correct and incorrect model specifications. We consider different scenarios, including omission of covariates, negligence of varying dispersion and misspecification of nonlinearity. Two applications using real data were performed.

Both the simulation results and applications yield important conclusions. When the mean response is scattered on the standard unit interval, the P^2 and R^2 coefficients perform similarly well, and both enable to identify usual model misspecification. On the other hand, it is noteworthy that when the response values are close to one or zero the P^2 criterion outperformed the R^2-like criteria in identifying problems on the model-fitting. Generally, these behaviors are related to influential observations and appropriated link functions for each range of response on standard unit interval. We notice that the log–log function models yield the best fits when the response is closer to zero, whereas the complementary log–log models yield the best fits when the response values are closer to one. These last conclusions were only supported by the P^2 criterion, but proved by the residual and influence analyses and by inference results. Another important conclusion is the poor performance of the R^2_{LR} criterion for beta regression models when the response is close to one of the standard unit interval boundaries. The R^2_{FC} outperforms the R^2_{LR} in identifying problems on the model variability on these ranges of the response variable. This conclusion is supported by the normal probability plots with simulated envelopes used in the real application.

Our proposed criterion proved to be very successful, since it selects the same models selected by the residual analysis, by the influence diagnostics and inference results. Despite this fact the normal probability plots with simulated envelopes reveal that questions about the model variability or the response distribution must be accessed the R^2-like criteria.

Therefore, to the class of beta regression models the best strategy to select the best model to fit a dataset is jointly used the P^2 and R^2_{FC} criteria. When the two criteria being simultaneously close to one, better shall be the fitted model.

Further work will be devoted to the theoretical properties of the P^2 statistic, and revisited statistical analysis, including post-Hoc analysis [26–28] with the Tukey's honestly significant difference test, and their p-values adjusted via false discovery rate [29] to highlight the existence of significant differences between the proposed and classical algorithms.

Author Contributions: Conceptualization, P.L.E.; Formal analysis, P.L.E., L.C.M.d.S., A.d.O.S. and R.O.; Investigation, P.L.E., L.C.M.d.S., A.d.O.S. and R.O.; Methodology, P.L.E., L.C.M.d.S., A.d.O.S. and R.O.; Software, P.L.E., L.C.M.d.S., A.d.O.S. and R.O.; Supervision, P.L.E.; Validation, P.L.E., L.C.M.d.S., A.d.O.S.; Visualization, P.L.E., L.C.M.d.S., A.d.O.S.; Writing—original draft, P.L.E. and R.O.; Writing—review & editing, P.L.E. and R.O.

Acknowledgments: This work was supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Ciência e Tecnologia de Pernambuco (FACEPE). The authors also thank three anonymous referees for comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Appendix

In what follows, we shall present the score function and Fisher's information for β and γ for the nonlinear beta regression models [1]. The log-likelihood function for the model (2) is given by $\ell(\beta,\gamma) = \sum_{t=1}^{n} \ell_t(\mu_t,\phi_t)$, and $\ell_t(\mu_t,\phi_t) = \log\Gamma(\phi_t) - \log\Gamma(\mu_t\phi_t) - \log\Gamma((1-\mu_t)\phi_t) + (\mu_t\phi_t - 1)\log y_t + \{(1-\mu_t)\phi_t - 1\}\log(1-y_t)$. The score function for β is

$$U_\beta(\beta,\gamma) = J_1^\top \Phi T(y^* - \mu^*), \tag{A1}$$

where $J_1 = \partial\eta_1/\partial\beta$ (an $n \times k$ matrix), $\Phi = \mathrm{diag}\{\phi_1,\ldots,\phi_n\}$, the tth elements of y^* and μ^* being given in (6). Also, $T = \mathrm{diag}\{1/g'(\mu_1),\ldots,1/g'(\mu_n)\}$. The score function for γ can be written as $U_\gamma(\beta,\gamma) = J_2^\top Ha$, where $J_2 = \partial\eta_2/\partial\gamma$ (an $n \times q$ matrix), the tth element of the vector a is $a_t = \mu_t(y_t^* - \mu_t^*) + \log(1-y_t) - \psi((1-\mu_t)\phi_t) + \psi(\phi_t), t = 1,\ldots,n$ and $H = \mathrm{diag}\{1/h'(\phi_1),\ldots,1/h'(\phi_n)\}$. The components of Fisher's information matrix are

$$K_{\beta\beta} = J_1^\top \Phi W J_1^\top, \quad K_{\beta\gamma} = K_{\gamma\beta}^\top = J_1^\top CTHJ_2^\top \quad \text{and} \quad K_{\gamma\gamma} = J_2^\top DJ_2^\top. \tag{A2}$$

Here, $W = \mathrm{diag}\{w_1,\ldots,w_n\}$, where

$$w_t = \phi_t v_t[1/\{g'(\mu_t)\}^2] \quad \text{and} \quad v_t = \{\psi'(\mu_t\phi_t) + \psi'((1-\mu_t)\phi_t)\}, \tag{A3}$$

$C = \mathrm{diag}\{c_1,\ldots,c_n\}$; $c_t = \phi_t\{\psi'(\mu_t\phi_t)\mu_t - \psi'((1-\mu_t)\phi_t)(1-\mu_t)\}$, $D = \mathrm{diag}\{d_1,\ldots,d_n\}$; $d_t = \xi_t/(h'(\mu_t))^2$ and $\xi_t = \{\psi'(\mu_t\phi_t)\mu_t^2 + \psi'((1-\mu_t)\phi_t)(1-\mu_t)^2 - \psi'(\phi_t)\}, 1,\ldots,n$

Local influence:

Let $\widehat{\theta}$ and $\widehat{\theta}_\delta$ be the ML estimators of θ for the assumed and perturbed models, respectively. The perturbation in the assumed model is introduced through a vector δ, $n \times 1$. The likelihood displacement $LD_\delta = 2\left\{\ell(\widehat{\theta}) - \ell(\widehat{\theta}_\delta)\right\}$ can be used to assess the influence of the perturbation on the ML estimate. Ref. [5] is interest to look for the direction $I_{\max}$, relative with a set of observations that corresponding to the largest likelihood displacement. The index plot of $I_{\max}$ can be used to single out observations that are jointly influential. Ref. [5] showed that $I_{\max}$ is the unit norm eigenvector corresponding to the largest eigenvalue

of $-\Delta^\top \ddot{\ell}^{-1}\Delta$. where $\ddot{\ell} = \partial^2\ell(\widehat{\theta})/\partial\theta\partial\theta^\top$ and Δ is a $s \times n$ matrix given by $\Delta = \partial^2\ell_\delta(\theta)/\partial\theta\partial\delta^\top$, evaluated at $\theta = \widehat{\theta}$ and $\delta = \delta_0$, which represents no perturbation.

On the other hand, the normal curvature in the direction of the tth individual, i.e., in the direction of the vector whose tth component equals one and all other elements are zero, becomes

$$C_t = 2|\Delta_t^\top \ddot{\ell}^{-1}\Delta_t|, \tag{A4}$$

where Δ_t is the tth column of Δ [19] C_t is the total local influence of observation t and observations such that $C_t > 2\sum_{t=1}^{n} C_t/n$ can be taken to be individually influential. We partition the parameter vector θ as $\theta = (\beta^\top, \gamma^\top)^\top$. Suppose we are interested in the local influence relative to β, then

$$C_{t;\beta} = 2|\Delta_t^\top(\ddot{\ell}^{-1} - \ddot{\ell}_{\gamma\gamma})\Delta_t| \quad \text{where} \quad \ddot{\ell}_{\gamma\gamma} = \frac{\partial^2\ell(\theta)}{\partial\gamma\partial\gamma^\top} \quad \text{and} \quad \ddot{\ell}_{\gamma\gamma} = \begin{pmatrix} 0 & 0 \\ 0 & \ddot{\ell}_{\gamma\gamma}^{-1} \end{pmatrix}. \tag{A5}$$

Similarly, the local influence relative to γ is given by

$$C_{t;\gamma} = 2|\Delta_t^\top(\ddot{\ell}^{-1} - \ddot{\ell}_{\beta\beta})\Delta_t| \quad \text{where} \quad \ddot{\ell}_{\beta\beta} = \frac{\partial^2\ell(\theta)}{\partial\beta\partial\beta^\top} \quad \text{and} \quad \ddot{\ell}_{\beta\beta} = \begin{pmatrix} \ddot{\ell}_{\beta\beta}^{-1} & 0 \\ 0 & 0 \end{pmatrix}. \tag{A6}$$

Here, the quantities $I_{\max;\beta}$ and $I_{\max;\gamma}$ are the unit norm eigenvector corresponding to the largest eigenvalue of $-\Delta^\top(\ddot{\ell}^{-1} - \ddot{\ell}_{\gamma\gamma})\Delta$ and $-\Delta^\top(\ddot{\ell}^{-1} - \ddot{\ell}_{\beta\beta})\Delta$, respectively. The most usual perturbation schemes are case-weight, response perturbation and covariate perturbation. Details of the Δ structure for each perturbation scheme and for the expression of $\ddot{\ell}^{-1}$ can be accessed by the local influence theory developed by [4] to the nonlinear beta regression models.

References

1. Simas, A.B.; Barreto-Souza, W.; Rocha, A.V. Improved estimators for a general class of beta regression models. *Comput. Stat. Data Anal.* **2010**, *54*, 348–366. [CrossRef]
2. Ospina, R.; Ferrari, S.L. Inflated beta distributions. *Stat. Pap.* **2010**, *51*, 111. [CrossRef]
3. Ospina, R.; Ferrari, S.L. A general class of zero-or-one inflated beta regression models. *Comput. Stat. Data Anal.* **2012**, *56*, 1609–1623. [CrossRef]
4. Rocha, A.V.; Simas, A.B. Influence diagnostics in a general class of beta regression models. *Test* **2011**, *20*, 95–119. [CrossRef]
5. Cook, R. Assessment of local influence (with discussion). *J. R. Stat. Soc. B* **1986**, *48*, 133–169.
6. Bayer, F.M.; Cribari-Neto, F. Model selection criteria in beta regression with varying dispersion. *Commun. Stat. Simul. Comput.* **2017**, *46*, 720–746. [CrossRef]
7. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*; Akadémiai Kiadó: Budapest, Hungary, 1973; pp. 267–281.
8. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
9. Allen, D.M. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **1974**, *16*, 125–127. [CrossRef]
10. Mediavilla, F.; Shah, V.A. A comparison of the coefficient of predictive power, the coefficient of determination and AIC for linear regression. *J. Appl. Bus. Econ.* **2008**, *8*, 44.
11. Özkale, M.R. Predictive performance of linear regression models. *Stat. Pap.* **2015**, *56*, 531–567. [CrossRef]
12. Cook, R.D. Detection of Influential Observation in Linear Regression. *Technometrics* **1977**, *19*, 15–18.
13. Ferrari, S.; Cribari-Neto, F. Beta regression for modelling rates and proportions. *J. Appl. Stat.* **2004**, *31*, 799–815. [CrossRef]

14. Pregibon, D. Logistic Regression Diagnostics. *Ann. Stat.* **1981**, *9*, 705–724, doi:10.1214/aos/1176345513. [CrossRef]
15. Espinheira, P.L.; Ferrari, S.L.; Cribari-Neto, F. On beta regression residuals. *J. Appl. Stat.* **2008**, *35*, 407–419. [CrossRef]
16. Walker, E.; Birch, J.B. Influence Measures in Ridge Regression. *Technometrics* **1988**, *30*, 221–227. [CrossRef]
17. Cook, R.D.; Weisberg, S. *Residuals and Influence in Regression*; Chapman and Hall: New York, NY, USA, 1982.
18. Espinheira, P.L.; Ferrari, S.L.P.; Cribari-Neto, F. Influence diagnostics in beta regression. *Comput. Stat. Data Anal.* **2008**, *52*, 4417–4431. [CrossRef]
19. Lesaffre, E.; Verbeke, G. Local Influence in Linear Mixed Models. *Biometrics* **1998**, *54*, 570–582. [CrossRef] [PubMed]
20. Nagelkerke, N. A note on a general definition of the coefficient of determination. *Biometrika* **1991**, *78*, 691–692. [CrossRef]
21. Doornik, J.A. *An Object-Oriented Matrix Programming Language Ox*, 6th ed.; Timberlake Consultants Ltd: London, UK, 2009.
22. Espinheira, P.L.; Silva, A.O. Nonlinear simplex regression models. *arXiv* **2018**, arXiv:1805.10843.
23. Espinheira, P.L.; Santos, E.G.; Cribari-Neto, F. On nonlinear beta regression residuals. *Biom. J.* **2017**, *59*, 445–461. [CrossRef] [PubMed]
24. Salazar, S.M.G. Contribuicion al Estudio de la Reaccion de Decomposicion de la Zeolita Y em Presencia de Vapor de Agua y Vanadio. Master's Thesis, Universidad Nacional de Colombia, Bogota, Colombia, 2005.
25. Espinheira, P.L.; Ferrari, S.L.; Cribari-Neto, F. Bootstrap prediction intervals in beta regressions. *Comput. Stat.* **2014**, *29*, 1263–1277. [CrossRef]
26. Rupert, G., Jr. *Simultaneous Statistical Inference*; Springer Science & Business Media: New York, NY, USA, 2012.
27. Triguero, I.; García, S.; Herrera, F. Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study. *Knowl. Inf. Syst.* **2015**, *42*, 245–284. [CrossRef]
28. Livieris, I.; Kanavos, A.; Tampakas, V.; Pintelas, P. An auto-adjustable semi-supervised self-training algorithm. *Algorithms* **2018**, *11*, 139. [CrossRef]
29. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **1995**, *57*, 289–300. [CrossRef]