Errors-in-variables beta regression models

Jalmar M. F. Carrasco[1], Silvia L. P. Ferrari[2,4] and Reinaldo B. Arellano-Valle[3]
[1] Departament of Statistics, Federal University of Bahia, Brazil,
[2] Departament of Statistics, University of São Paulo, Brazil
[3] Departament of Statistics, Pontifical Catholic University of Chile, Chile
[4] Corresponding author: Silvia L.P. Ferrari, e-mail:silviaferrari.usp@gmail.com

Abstract

Beta regression models provide an adequate approach for modeling continuous outcomes limited to the interval $(0, 1)$. This paper deals with an extension of beta regression models that allow for explanatory variables to be measured with error. The structural approach, in which the covariates measured with error are assumed to be random variables, is employed. Three estimation methods are presented, namely maximum likelihood, maximum pseudo-likelihood and regression calibration. Monte Carlo simulations are used to evaluate the performance of the proposed estimators and the naïve estimator. Also, a residual analysis for beta regression models with measurement errors is proposed. The results are illustrated in a real data set.

Keywords: Beta regression model; Errors-in-variables model; Gauss-Hermite quadrature; Maximum likelihood; Maximum pseudo-likelihood; Regression calibration.

1 Introduction

The beta regression models provide an adequate approach for modeling continuous outcomes limited to the interval $(0, 1)$, or more generally, limited to any open interval (a, b) as long as the limits are known (Ferrari and Cribari-Neto, 2004). Although the literature on beta regression has grown fast in the last few years, errors-in-variables models with beta distributed outcomes is an unexplored area.

A beta regression model assumes that the response variable, y, has a beta distribution with probability density function

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

where $\Gamma(\cdot)$ is the gamma function, $0 < \mu < 1$ and $\phi > 0$, and we write $y \sim \mathrm{Beta}(\mu, \phi)$. Here, $\mu = \mathrm{E}(y)$ and ϕ is regarded as a precision parameter since $\mathrm{Var}(y) = \mu(1-\mu)/(1+\phi)$. For independent observations $y_1, y_2, \ldots, y_n$, where each y_t follows a beta density (1) with mean μ_t and unknown precision parameter ϕ, the beta regression model defined by Ferrari and Cribari-Neto (2004) assumes that

$$g(\mu_t) = \mathbf{z}_t^\top \boldsymbol{\alpha}, \tag{2}$$

with $\boldsymbol{\alpha} \in \mathbb{R}^{p_\alpha}$ being a column vector of unknown parameters, and with $\mathbf{z}_t^\top = (z_{t1}, \ldots, z_{tp_\alpha})$ being a vector of p_α fixed covariates $(p_\alpha < n)$. The link function $g(\cdot) : (0, 1) \to \mathbb{R}$ is assumed to be a continuous, strictly monotone and twice differentiable function. There are many possible choices for $g(\cdot)$, for instance, the logit link, $g(\mu_t) = \log[\mu_t/(1 - \mu_t)]$, the probit link, $g(\mu_t) = \Phi^{-1}(\mu_t)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and the complementary log-log link, $g(\mu_t) = \log[-\log(1 - \mu_t)]$.

Extensions for the beta regression model proposed by Ferrari and Cribari-Neto (2004) that allow the precision parameter to vary across observations, or that involve non-linear structures for the regression specification of the mean and the precision parameter, are presented by Smithson and Verkuilen (2006), Simas *et al.* (2010), among others. The beta regression model with linear specification for the transformed mean and precision parameter is given by (1), (2) and

$$h(\phi_t) \quad = \quad \mathbf{v}_t^\top \boldsymbol{\gamma}, \tag{3}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^{p_\gamma}$ $(p_\alpha + p_\gamma < n)$ is a column vector of unknown parameters, $\mathbf{v}_t = (v_{t1}, \cdots, v_{tp_\gamma})^\top$ is a vector of fixed covariates, $h(\cdot) : (0, \infty) \longrightarrow \mathbb{R}$ is a strictly monotone, twice differentiable link function. A possible choice for $h(\cdot)$ is $h(\phi_t) = \log(\phi_t)$.

The purpose of this paper is to extend the beta regression model (1)-(3) to the situation where some covariates are not directly measured or are measured with error. A practical application of errors-in-variables beta regression models will be illustrated in a study of the risk of coronary heart disease as a function of low-density lipoprotein (LDL) cholesterol level ("bad cholesterol") and body mass index (BMI). The dataset consists of observations of systolic blood pressure (SBP), diastolic blood pressure (DBP), BMI and total cholesterol level (TC) in a group of 182 smoking women aged 50 to 87 years. The total cholesterol may be considered as a surrogate of LDL, which is a covariate of interest, and whose direct measure is more expensive and time consuming. The difference between SBP and DBP results in what is known as the pulse pressure, $PP = SBP - DBP$, and the relative pulse pressure is $RPP = (SBP - DBP)/SBP = PP/SBP$. Small values of RPP, $RPP < 25\%$ say, is indicative of risk of heart disease (American College of Surgeons, 2008, p. 58). Notice that the response variable, RPP, is continuous and limited to the unit interval, and that one of the covariates, namely LDL, is not measured directly.

2 Model and likelihood

Let $y_1, \ldots, y_n$ be independent observable random variables arising from a sample of size n, such that y_t has a beta distribution with probability density function (1) with parameters $\mu = \mu_t$ and $\phi = \phi_t$. In the following, we assume that μ_t and ϕ_t may depend on covariates and unknown parameters. In practice, some covariates may not be precisely observed, but, instead, may be obtained with error. The model considered in this paper assumes a linear structure for the specification of the mean and the precision parameters, and also assumes that both specifications may involve covariates measured with error. Specifically, we replace the mean submodel (2) and the precision submodel (3) by

$$\begin{aligned} g(\mu_t) \quad &= \quad \mathbf{z}_t^\top \boldsymbol{\alpha} + \mathbf{x}_t^\top \boldsymbol{\beta}, &(4)\\ h(\phi_t) \quad &= \quad \mathbf{v}_t^\top \boldsymbol{\gamma} + \mathbf{m}_t^\top \boldsymbol{\lambda}, &(5)\end{aligned}$$

respectively, where $\boldsymbol{\beta} \in \mathbb{R}^{p_\beta}$, $\boldsymbol{\lambda} \in \mathbb{R}^{p_\lambda}$ are column vectors of unknown parameters, $\mathbf{x}_t = (x_{t1}, \cdots, x_{tp_\beta})^\top$ and $\mathbf{m}_t = (m_{t1}, \cdots, m_{tp_\lambda})^\top$ $(p_\alpha + p_\beta + p_\gamma + p_\lambda < n)$ are unobservable (latent) covariates, in the sense that they are observed with error. The vectors of covariates measured without error, $\mathbf{z}_t$ and $\mathbf{v}_t$, may contain variables in common, and likewise, $\mathbf{x}_t$ and $\mathbf{m}_t$. Let $\mathbf{s}_t$ be the vector containing all the unobservable covariates. For $t = 1, \ldots, n$, the random vector $\mathbf{w}_t$ is observed in place of $\mathbf{s}_t$, and it is assumed that

$$\mathbf{w}_t = \boldsymbol{\tau}_0 + \boldsymbol{\tau}_1 \circ \mathbf{s}_t + \mathbf{e}_t, \tag{6}$$

where $\mathbf{e}_t$ is a vector of random errors, $\boldsymbol{\tau}_0$ and $\boldsymbol{\tau}_1$ are (possibly unknown) parameter vectors and $\circ$ represents the Hadamard (elementwise) product. The parameter vectors $\boldsymbol{\tau}_0$ and $\boldsymbol{\tau}_1$ can be interpreted as the additive and multiplicative biases of the measurement error mechanism, respectively. If $\boldsymbol{\tau}_0$ is a vector of zeros and $\boldsymbol{\tau}_1$ is a vector of ones, we have the classical additive model $\mathbf{w}_t = \mathbf{s}_t + \mathbf{e}_t$. Here, we follow the structural approach, in which the unobservable covariates are regarded as random variables, i.e. we assume that $\mathbf{s}_1, \ldots, \mathbf{s}_n$ are independent and identically distributed random vectors. In this case, it is also usual to assume that they are independent of the measurement errors $\mathbf{e}_1, \ldots, \mathbf{e}_n$. Moreover, the normality assumption for the joint distribution of $\mathbf{s}_t$ and $\mathbf{e}_t$ is assumed. The parameters of the joint distribution of $\mathbf{w}_t$ and $\mathbf{s}_t$ is denoted by $\boldsymbol{\delta}$.

Let $(y_1, \mathbf{w}_1), \ldots, (y_n, \mathbf{w}_n)$ be the observable variables. We omit the observable vectors $\mathbf{z}_t$ and $\mathbf{v}_t$ in the notation as they are non-random and known. The joint density function of $(y_t, \mathbf{w}_t)$, which is the observation for the t-th individual, is obtained by integrating the joint density of the complete data $(y_t, \mathbf{w}_t, \mathbf{s}_t)$,

$$f(y_t, \mathbf{w}_t, \mathbf{s}_t; \boldsymbol{\theta}, \boldsymbol{\delta}) = f(y_t | \mathbf{w_t}, \mathbf{s}_t; \boldsymbol{\theta}) f(\mathbf{s}_t, \mathbf{w}_t; \boldsymbol{\delta}),$$

with respect to $\mathbf{s}_t$. Here, $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\lambda}^\top)^\top$ represents the parameter of interest, and $\boldsymbol{\delta}$ is the nuisance parameter. The joint density $f(\mathbf{w}_t, \mathbf{s}_t; \boldsymbol{\delta})$, which is associated to the measurement error model, can be written as $f(\mathbf{w}_t, \mathbf{s}_t; \boldsymbol{\delta}) = f(\mathbf{w}_t | \mathbf{s}_t; \boldsymbol{\delta}) f(\mathbf{s}_t | \boldsymbol{\delta})$ as well as $f(\mathbf{w}_t, \mathbf{s}_t; \boldsymbol{\delta}) = f(\mathbf{s}_t | \mathbf{w}_t; \boldsymbol{\delta}) f(\mathbf{w}_t | \boldsymbol{\delta})$. In this work we assume that, given the true (unobservable) covariates $\mathbf{s}_t$, the response variable y_t does not depend on the surrogate covariates $\mathbf{w}_t$; i.e. $f(y_t | \mathbf{w}_t, \mathbf{s}_t; \boldsymbol{\theta}) = f(y_t | \mathbf{s}_t; \boldsymbol{\theta})$. In other words, conditionally on $\mathbf{s}_t$, y_t and $\mathbf{w}_t$ are assumed to be independent (Bolfarine and Arellano-Valle, 1998). Therefore, the log-likelihood function for a sample of n observations is given by

$$\begin{aligned}
\ell(\boldsymbol{\theta}, \boldsymbol{\delta}) &= \sum_{t=1}^{n} \log f(\mathbf{w}_t; \boldsymbol{\delta}) + \sum_{i=1}^{n} \log \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(y_t | \mathbf{s}_t; \boldsymbol{\theta}) f(\mathbf{s}_t | \mathbf{w}_t; \boldsymbol{\delta}) d\mathbf{s}_t \\
&= \sum_{t=1}^{n} \ell_{1t}(\boldsymbol{\delta}) + \sum_{t=1}^{n} \ell_{2t}(\boldsymbol{\theta}, \boldsymbol{\delta}).
\end{aligned} \tag{7}$$

In general, the likelihood function involves analytically intractable integrals and, hence, approximate inference methods need to be considered. In the next section, we present three different approaches to estimate the parameters.

3 Estimation

The second term of the log-likelihood function $\ell(\boldsymbol{\theta}, \boldsymbol{\delta})$ in (7), which depends on a non-analytical integral, can be approximated using the Gauss-Hermite quadrature; see, for instance, Abramowitz and Stegun (1972, Chapter 22).

We present three different strategies to estimate the parameters $\boldsymbol{\theta}$. The first one is the maximization of an approximate log-likelihood function obtained by approximating the second term of (7) using the Gauss-Hermite quadrature. The second one is a two-step procedure (Guolo, 2011). First, the nuisance parameter vector $\boldsymbol{\delta}$ is estimated by maximizing the reduced log-likelihood function

$$\ell_r(\boldsymbol{\delta}) = \sum_{t=1}^{n} \ell_{1t}(\boldsymbol{\delta}). \tag{8}$$

Second, the estimate $\widehat{\boldsymbol{\delta}}$ obtained from the maximization of (8) is inserted in the original log-likelihood function (7), which results in the pseudo-log-likelihood function

$$\ell_p(\boldsymbol{\theta};\widehat{\boldsymbol{\delta}}) = \sum_{t=1}^{n} \ell_{1t}(\widehat{\boldsymbol{\delta}}) + \sum_{t=1}^{n} \ell_{2t}(\boldsymbol{\theta},\widehat{\boldsymbol{\delta}}). \tag{9}$$

As in $\ell(\boldsymbol{\theta},\boldsymbol{\delta})$, the second term in $\ell_p(\boldsymbol{\theta};\widehat{\boldsymbol{\delta}})$ cannot be expressed in closed form and requires numerical integration. However, unlike the integral in $\ell(\boldsymbol{\theta},\boldsymbol{\delta})$, the integral in $\ell_p(\boldsymbol{\theta};\widehat{\boldsymbol{\delta}})$ depends on the parameter of interest only. It is possible to approximate $\ell_{2t}(\boldsymbol{\theta},\widehat{\boldsymbol{\delta}})$ by a summation using the Gauss-Hermite quadrature.

The third strategy is the regression calibration method; see Carroll *et al.* (2006, Chap. 4), Freedman *et al.* (2008), Thurston *et al.* (2005) and Guolo (2011). The central idea is to replace the unobservable variable by an estimate of its conditional expected value given the observed covariates in the likelihood function. It is well known that regression calibration estimators are, in general, inconsistent. Skrondal and Kuha (2012) point out that “the inconsistency is typically small when the true effects of the covariates measured with error are moderate and/or the measurement error variance are small, but more pronounced when these conditions do not hold.”

Numerical properties of the three estimators described above are presented in the full version of this paper. Also, diagnostic tools for error-in-variables beta regression models are presented and the dataset mentioned in Section 1 is analysed using the proposed model and inferential methods.

4 Concluding remarks

In this paper we proposed and studied errors-in-variables beta regression models. We proposed three different estimation methods, namely, the approximate maximum likelihood, maximum pseudo-likelihood and regression calibration methods. We performed a Monte Carlo simulation study to compare the performance of the estimators in terms of bias, root-mean-square errors and coverage of confidence intervals. Overall, we reached the following conclusions. First, ignoring the measurement error may lead to severely biased inference. Second, the regression calibration approach is very simple and seems to be reliable for estimating the parameters of the mean submodel when the measurement error variance is small. However, there is clear indication that it is not consistent for estimating the parameters that model the precision of the data.Third, the approximate maximum likelihood and maximum pseudo-likelihood approaches perform well, the later being less computationally demanding than the former. We, therefore, recommend the maximum pseudo-likelihood estimation for practical applications. We emphasize that the maximum pseudo-likelihood estimator coincides with the improved regression calibration estimator proposed by Skrondal and Kuha (2012). Its consistency and asymptotic normality are justified by these authors. We also proposed a standardized weighted residual for diagnostic purposes. All our results were illustrated in the analysis of a real data set.

Acknowledgements

We gratefully acknowledge the financial support from CAPES-Brazil, CNPq-Brazil, FAPESP-Brazil and FONDECYT (1120121)-Chile.

References

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions*. New York:Dover.

American College of Surgeons (2008). *ATLS Advanced Trauma Life Support Program for Doctors*. Chicago: American College of Surgeons.

Bolfarine, H. and Arellano-Valle, R. B. (1998). Weak nondifferential measurement error models. *Statistics and Probability Letters*, **40**, 279–287.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. New York: Chapman and Hall.

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

Freedman, L. S., Midthune, D., Carroll, R., and Kipnis, V. (2008). A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Statistics in Medicine*, **27**, 5195–5216.

Guolo, A. (2011). Pseudo-likelihood inference for regression models with misclassified and mismeasured variables. *Statistica Sinica*, **21**, 1639–1663.

Simas, A. B., Barreto-Souza, W., and Rocha, A. V. (2010). Improved estimators for a general class of beta regression models. *Computational Statistics and Data Analysis*, **54**, 348–366.

Skrondal, A. and Kuha, J. (2012). Improved regression calibration. *Psychometrika*, **77**, 649–669.

Smithson, M. and Verkuilen, J. (2006). A better lemon-squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, **11**, 54–71.

Thurston, S. W., Williams, P. L., Hauser, R., Hu, H., Hernandez-Avila, M., and Spiegelman, D. (2005). A comparison of regression calibration approaches for designs with internal validation data. *Journal of Statistical Planning and Inference*, **131**, 175–190.