Multivariate beta regression with application to small area estimation

Debora Ferreira de Souza
debora@dme.ufrj.br
Fernando Antônio da Silva Moura
fmoura@im.ufrj.br
Departamento de Métodos Estatísticos - UFRJ

Abstract:

Multivariate beta regression models for jointly modeling two or more variables whose values belong to the $(0, 1)$ interval, such as indexes, rates or proportions are proposed. The multivariate model can help the estimation process by borrowing information between units and obtaining more precise estimates, especially for small samples. Each response variable was assumed to be beta distributed, allowing to deal with multivariate asymmetric data.

Copula functions are used to construct the joint distribution of the dependent variables, with all the marginal distributions fixed as beta. A hierarchical beta regression model is also proposed with correlated random effects. Both models have shown to be useful for making small area prediction. The inference process was conducted using full Bayesian approach.

We present an application for estimating two index of educational attainment at school and municipality levels of a Brazilian state. Our predictions are compared with others approaches commonly employed in practice.

Key-words: Univariate beta regression, MCMC, missing values prediction and small domain, education evaluation.

1 Introduction

We propose a new approach to jointly modeling indexes, rates or proportions, commonly estimated with low accuracy in small samples. Examples of variables measured in the range $(0, 1)$ and related to each other are the proportion of poor people, mortality rate, the ratio of expense with food to total expense. While the motivation of this work has been the estimation of rates or proportions in small areas (or domains), the strategy used to achieve this goal can be applied to a more general context.

Multivariate models are developed for modeling rates or proportions, offering the possibility of jointly dealing with related quantities in one single model and enjoying the benefits that this joint approach offers. The exchange of information between variables in the multivariate models proposed here can help to obtain more precise estimates of the quantities of interest.

In recent years, numerous applications involving the Beta distribution has been developed due to its appropriateness for modeling rates or proportions, it is defined in the range $(0, 1)$, it allows for asymmetry present in these types of variables and it assumes different forms depending on their parameters. The beta regression also allows heteroscedastic observations. However, the most proposed use of Beta distribution in the context of regression has been restricted to cases where there is only one dependent variable. Furthermore, under the Bayesian approach, there are few works and applications involving beta regression models.

This paper develops multivariate regression models where the dependent variables marginally follow a Beta distribution. These models were developed to address data fitting in general contexts, but application in small area estimation shows that they are especially advantageous in this situation. In the models proposed throughout this paper, the response variables do not add one, as in some models to proportions discussed, for example, in Melo et al. (2009) and Fabrizi et al. (2011).

The proposed multivariate models assume that the Beta marginal distributions were reparametrized by the mean and the dispersion, as in Ferrari and Cribari-Neto (2004). The association between the response variables is considered through a copula function applied to the marginal densities. Copulas are useful tools for building multivariate distributions where the marginal distributions are given or known, allowing individual models be analyzed together. Additionally, they allow the representation of various types of dependence between variables. Copulation allows flexibility for handling non-linear

relationships between the response variables, and therefore is a more general setup than the Multivariate Normal distribution, which allows only linear relationships.

Basically, two classes of multivariate models with Beta response are proposed: a beta regression model, where the marginal are connected by a copula function and a hierarchical beta model with correlation between their means. Both models can be used in situations where the researcher needs to jointly analyze data from related response variables in the range $(0, 1)$. They can be used to improve prediction of observations and target population parameters in small areas estimation.

In small area estimation context where there are auxiliary variables and data from multiple characteristics available, it is possible to propose and apply a multivariate model. Several authors argue that this approach provides better estimates, because takes into account the the correlations between the response variables left after conditioning on the auxiliary variables. Fay (1987) proposed to model the joint behavior of the median income in households of three, four and five dwellers. Datta et al. (1999) applied a multivariate mixed linear model and concluded from a simulation study that the multivariate approach provides better results than setting a separate model for each variable. For example, the methods most commonly employed are based on borrowing information between neighbor areas or related ones. The models proposed in this paper have direct application to the small area estimation problem by providing the exchange of information between the response variables.

The article is organized as follows. In Section 2 we propose a Multivariate Beta Regression model by employing copula functions. In Section 3 we applied our proposed models to the small area estimation problems, presenting two applications to Brazilian education data. Section 4 offers some conclusions and suggestions for further research.

2 Multivariate Beta Regression model based on copulas

The structure of dependence between two or more related response variables can be defined in terms of their joint distribution. One way of obtaining a multivariate beta distribution is to join the univariate beta through copulation, which is one of the most useful tools when the marginal distributions are given or known. The use of copula functions enables the representation of various types of dependence between variables. In practice, this implies a more flexible assumptions about the form of the joint distribution than that given in Olkin and Liu (2003), which assumes that the marginal distributions have the

same parameter. For a complete study about copula function and its utilities in statistics, see Nelsen (2006).

Let $Y_1, ..., Y_K$ be functions of random variables with marginal distributions $F_1, ..., F_K$, respectively, and joint cumulative distribution function $H(y_1, ..., y_K) = C(F_1(y_1), ..., F_K(y_K))$, where $F_i \sim U(0,1)$, $i = 1, ..., K$ and $C(\cdot)$ is a copula function. Then the density function of $(Y_1, ..., Y_K)$ is given by:

$$\begin{aligned} h(y_1, ..., y_K) &= \frac{\partial^n H(y_1, ..., y_K)}{\partial y_1, ..., \partial y_K} \\ &= \frac{\partial^n C(F_1(y_1), ..., F_K(y_K))}{\partial F_1(y_1), ..., \partial F_K(y_K)} \times \frac{\partial F_1(y_1)}{\partial y_1} \times \cdots \times \frac{\partial F_K(y_K)}{\partial y_K} \\ &= c(F_1(y_1), ..., F_K(y_K)) \prod_{i=j}^{K} f_j(y_j) \end{aligned} \tag{1}$$

where

$$c(F_1(y_1), ..., F_K(y_K)) = \frac{\partial^n C(F_1(y_1), ..., F_K(y_K))}{\partial F_1(y_1), ..., \partial F_K(y_K)} \quad \text{and} \quad f_j(y_j) = \frac{\partial F_j(y_j)}{\partial y_j}, j = 1, ..., K.$$

Let $\mathbf{y} = ((y_{11}, ..., y_{K1}), ..., (y_{1n}, ..., y_{Kn}))$ be a random sample of size n from the density in (1). Thus, the likelihood function is given by:

$$L(\mathbf{\Psi}) = \prod_{i=1}^{n} c(F_1(y_{1i}|\mathbf{\Psi}), ..., F_K(y_{Ki}|\mathbf{\Psi})) f_1(y_{1i}|\mathbf{\Psi}) ... f_K(y_{Ki}|\mathbf{\Psi})$$

where $\mathbf{\Psi}$ denotes the set of parameters that define the cumulative distribution functions and the density, as well as the copula function.

We assume that each response variable is beta distributed and the structure of dependence between them is defined by their joint distribution which is obtained by applying a copula function. Thus, the multivariate regression model proposed is such that:

$$\begin{aligned} y_{ij}|\mu_{ij}, \phi_j &\sim Be(\mu_{ij}, \phi_j), \quad i = 1, ..., n, \quad j = 1, ..., K \\ g(\mu_{ij}) &= \eta_{ij} = \sum_{l=1}^{p} x_{il}\beta_{lj} \end{aligned}$$

where $g(\cdot)$ is the link function. Under the Bayesian approach, the specification of the model is completed by assigning a prior distribution to $\boldsymbol{\phi} = (\phi_1, ..., \phi_K)$,

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1K} \\ \vdots & \vdots & \vdots \\ \beta_{p1} & \cdots & \beta_{pK} \end{pmatrix}$$

and to the parameters that define the copula family.

The linear correlation coefficient is not suitable to measure the dependence between variables in a model involving copulation. One most appropriate measure, which can be found in Nelsen (2006), is the statistic τ of Kendall, given by

$$\tau = 4 \int_0^1 \int_0^1 C(u,v) dC(u,v) - 1.$$

The measure τ of Kendall is related to the parameter θ and can be used to assign a prior to θ. In this work, we focus on the bivariate case.

2.1 Multivariate hierarchical beta regression model

In the previous section was presented a multivariate beta regression model in which the marginal Beta regression coefficients regression were fixed. However, there are situations where it makes sense to assume that some or all of the coefficients are random. In these cases, the coefficients of each observation has a common average, suffering the influence of non-observable effects. Such models are often called mixed effects models with response in the exponential family, with applications in several areas. Jiang (2007) discusses linear mixed models and some inference procedures for estimating its parameters. Rao (2003) shows some use of mixed effects models in estimation in small areas.

In this section we propose a generalization of the multivariate regression model presented in Section 2 by assuming that some or all of the coefficients associated with the linear predictor of each response variable can be random and correlated.

Let y_{ijk} be the observed value of the k^{th} response variable in the j^{th} unit of the i^{th} area, $k = 1, ..., K$, $j = 1, ..., N_i$ and $i = 1, ..., M$. Furthermore, let us assume that y_{ijk} and $y_{i'jk}$ are conditional independents, $\forall\ i \neq i'$. The multivariate hierarchical beta regression model is defined as:

$$\begin{aligned}
\mathbf{y}_{ij} &\sim BetaM(\boldsymbol{\mu}_{ij}, \boldsymbol{\phi}, \boldsymbol{\theta}), \quad j = 1, ..., N_i, \quad i = 1, ..., M && (2)\\
g(\mu_{ijk}) &= \mathbf{x}_{ij}^T \boldsymbol{\lambda}_{ik}, \quad k = 1, ..., K && (3)\\
\lambda_{ilk} &= \beta_{lk} + \nu_{ilk}, && (4)\\
\boldsymbol{\nu}_{il} &= (\nu_{il1}, ..., \nu_{ilK})^T \sim N_K(\mathbf{0}, \boldsymbol{\Sigma}_l), \quad l = 1, ..., p && (5)
\end{aligned}$$

where: $BetaM(\boldsymbol{\mu}_{ij}, \boldsymbol{\phi}, \boldsymbol{\theta})$ denotes an Beta multivariate distribution built by using a copula function with parameter $\boldsymbol{\theta}$ and the Beta marginal distributions; $\mathbf{y}_{ij} = (y_{ij1}, ..., y_{ijK})$;

$\mathbf{x}_{ij}^T = (x_{ij1}, ..., x_{ijp})$; $\boldsymbol{\lambda}_{ik} = (\lambda_{i1k}, ..., \lambda_{ipk})$; $\boldsymbol{\phi} = (\phi_1, ..., \phi_K)$;

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1K} \\ \vdots & \cdots & \vdots \\ \beta_{p1} & \cdots & \beta_{pK} \end{pmatrix}; \quad \text{and} \quad \mathbf{x}_i^T = \begin{pmatrix} x_{i11} & \cdots & x_{i1p} \\ x_{i21} & \cdots & x_{i2p} \\ \vdots & \cdots & \vdots \\ x_{iM_i1} & \cdots & x_{iM_ip} \end{pmatrix}.$$

From (4) and (5) follows $\lambda_{ilk} \sim N(\beta_{lk}, \sigma_{lk}^2)$, $l = 1, ..., p$, $k = 1, ..., K$.

As generally described in equations (2), (3) and (4), the model allows all regression coefficients to be random, however, in many applications of hierarchical models only some coefficients are assumed to be random, specially the intercept term. In the model (2)-(5) all random effects in $\boldsymbol{\nu}$ could be considered independent and only the correlation between the response variables would be contemplated. However, to allow the averages of the responses also exchange information among themselves, it is considered that within each level i, and for each coefficient of the response variable l, the random effects concerning the response variables are correlated, i.e: $\boldsymbol{\nu}_{il} = (\nu_{il1}, ..., \nu_{ilK})^T \sim N_K(\mathbf{0}, \boldsymbol{\Sigma}_l)$ where

$$\boldsymbol{\Sigma}_l = \begin{pmatrix} \sigma_{l1}^2 & \sigma_{l12} & \cdots & \sigma_{l1K} \\ \sigma_{l12} & \sigma_{l2}^2 & \cdots & \sigma_{l2K} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{l1K} & \sigma_{l2K} & \cdots & \sigma_{lK}^2 \end{pmatrix}.$$

In this model, the dependence of the response variables appears at two levels: at the observations and at the linear predictors. This can be a point in favor of this model with respect to the small area estimation problem, because it allows the exchange of information between the means, which are interpreted as the true values of indices, rates or proportions of interest. The logistic link function was used in all applications. The model (2)-(5) assumes that information about K response variables and M areas, with m_i units, $i = 1, ..., M$ are available.

The equation (3) relates the averages of the response variables in each i^{th} area, and considers specific area effects. Thus, the mean μ_{ijk} and $\mu_{ijk'}$ also exchange information among themselves due to the fact that they are correlated. This exchange is particularly important in the small area estimation problem in which μ_{ijk} is interpreted as the true value of the rate or proportion of interest and information from related quantities can produce more accurate estimators.

3 Application to the small area estimation

The models defined in Section 2 were developed for being applied in general applications where there are K related variables, measured in the range $(0, 1)$, or an interval (a, b), which can be explained by p covariates. However, they may be also applied to the small area estimation problem.

Missing values can be easily treated as unknown parameters and included in the posterior distribution for being estimated when there is no observations in all the response variables, or in just one of them. As far as the small area estimation problem is concerned, the researcher may be interested in the estimation of functions of the response variables for units and/or areas not selected in the sample. Even in those areas where there is some information, the sample size may be small for the direct estimator provide reliable estimates with acceptable accuracy. The multivariate models proposed in the previous section can be applied for making predictions on the non-selected areas and for producing more accurate estimates for the selected areas. The missing values here are produced by the sampling design, and treated as unknown parameters to be estimated by the models. Auxiliary information (covariates) must be known for all units of the level to which one wants to make prediction. They can be obtained from a census or administrative records. We have not considered missing values in explanatory variables. The lower is the loss of data, the greater is the efficiency of the estimation. However, in a survey sampling where the main objective is to provide information to a higher level, the most common scenario is to have few selected areas, making it even difficult the estimation process.

In the following sections are presented two small area estimation application where predictions are made for non-selected areas. For both applications, the predictions for the small areas and for aggregation of them are provided.

3.1 Brazilian educational data

The Brazilian evaluation of the basic education is carried out by the Brazilian National of Education Research (INEP). It aims to evaluate the performance of students from the 4^{th} to the 8^{th} series of the elementary school. The tests are applied every two years to urban public schools with more than 20 students. The evaluation of Brazilian education combines performance on the reading and mathematics tests with socioeconomic information.

The hierarchical structure of the data, organized into municipalities, schools and

students, suggested the use of hierarchical modeling. Because the tests are applied to the entire universe of schools and students, it is possible to obtain the true quantities of interest. Therefore, it is possible to compare the estimates provided by the multivariate hierarchical beta model with the true observed values.

The multivariate hierarchical beta model is applied to two different situations, derived from two sampling procedures. Only schools with students of the 4^{th} series belong to the municipal administration in the Rio de Janeiro state were considered.

3.2 Application 1

In this application, for each municipality of Rio de Janeiro state was selected a sample of schools with equal probabilities and all students who took the tests were selected. Therefore, for each selected school, the proportions of correct answers in Portuguese and Mathematics are not estimates, but actual values.

The aim for this application is to estimate the proportions of correct answers in Portuguese Language and Mathematics for non-sampled schools. Let y_{ijk} be the average proportion of corrected answers observed in k^{th} subject for the j^{th} selected school in the i^{th} municipality and suppose that $y_{ijk} \sim Beta(\mu_{ijk}, \phi_k)$. The assumption that the true value follows a distribution is reasonable because y_{ijk} represents the proportion obtained in a single test. However, if it were applied others equivalent tests, we would expect that the average of corrected answers was μ_{ijk}. Thus, it justifies to modeling the observed average school y_{ijk} by a probability distribution.

The information available about the characteristics of school are provided by the questionnaires applied to schools directors and teachers. Schools where there were no answers for at least one of these questionnaires were excluded from the analysis. Municipalities where there was only one public school, after the first mentioned dropout were also eliminated, leaving 82 municipalities. For each one of these 82 municipalities, random sample of 20% of the schools were selected, with a minimum sample size of two schools per municipality. In 11 municipalities, all schools were selected. From the total of 1787 schools belong to the 82 municipalities, only 421 were selected.

It is assumed that there is information for all schools on the following chosen covariates: existence of the program to avoid school abandonment (x_2) and the percentage of teachers who teach less than 60% of the program of their disciplines (x_3).

The multivariate beta hierarchical model with and without a copula function were fitted

to the data. In each selected schools, all students were investigated, and the proportion of correct answers in each school correspond to the value observed by the survey. For both models, we have $M = 82$ municipalities. Let denote by N_i and n_i, the total number of schools and the number of schools sampled in the municipality i, thus $\sum_{i=1}^{M} N_i = 1787$ and $\sum_{i=1}^{M} m_i = 421$. Only the intercepts were considered random. The fitted models were:
Model 1:

$$\begin{aligned} y_{ijk} &\sim Beta(\mu_{ijk}, \phi_k), \quad j = 1, ..., M_i, \quad i = 1, ..., M \\ g(\mu_{ijk}) &= \lambda_{i1k} + x_{ij2}\beta_{2k} + x_{ij3}\beta_{3k}, \quad k = 1, 2 \\ \lambda_{i1k} &= \beta_{1k} + \nu_{i1k} \\ (\nu_{i11}, \nu_{i12})^T &\sim N_2(\mathbf{0}, \mathbf{\Sigma}). \end{aligned}$$

Model 2:

$$\begin{aligned} \mathbf{y}_{ij} &\sim BetaM(\boldsymbol{\mu}_{ij}, \boldsymbol{\phi}, \theta), \quad j = 1, ..., M_i, \quad i = 1, ..., M \\ g(\mu_{ijk}) &= \lambda_{i1k} + x_{ij2}\beta_{2k} + x_{ij3}\beta_{3k}, \quad k = 1, 2 \\ \lambda_{i1k} &= \beta_{1k} + \nu_{i1k} \\ (\nu_{i11}, \nu_{i12})^T &\sim N_2(\mathbf{0}, \mathbf{\Sigma}), \end{aligned}$$

where $BetaM(\boldsymbol{\mu}_{ij}, \boldsymbol{\phi}, \theta)$ denotes a Beta bivariate distribution built by employing a Gaussian copula with parameter θ and beta marginal distributions parametrized by $\boldsymbol{\mu}_{ij} = (\mu_{ij1}, \mu_{ij2})^T$ and $\boldsymbol{\phi} = (\phi_1, \phi_2)$. The Gaussian copula were use in this application because is quite flexible, since the statistics τ of Kendall is on the interval $[-1, 1]$ and it is so applied in many situations.

3.2.1 Inference

For both models, it is assumed that the population model holds for the sample, i.e., sample selection bias is absent, see Pfeffermann et al. (2006).

Let $\mathbf{y}_o$ and $\mathbf{y}_f$ be, the matrices of the response variables for the sampled and non-sampled schools, respectively, and $\mathbf{W} = \mathbf{\Sigma}^{-1}$. The posterior density for the model 2 of all unknown quantities, including $\mathbf{y}_f$ is given by:

$$\begin{aligned} p(\mathbf{y}_f, \boldsymbol{\beta}, \boldsymbol{\phi}, \theta, \boldsymbol{\lambda}, \mathbf{W} | \mathbf{y}_o) &\propto p(\mathbf{y}_o | \boldsymbol{\beta}, \boldsymbol{\phi}, \theta, \boldsymbol{\lambda}, \mathbf{W}, \mathbf{y}_f) p(\mathbf{y}_f | \boldsymbol{\beta}, \boldsymbol{\phi}, \theta, \boldsymbol{\lambda}, \mathbf{W}) \\ &\times p(\boldsymbol{\lambda} | \boldsymbol{\beta}, \boldsymbol{\phi}, \theta, \mathbf{W}) p(\boldsymbol{\beta}) p(\boldsymbol{\phi}) p(\theta) p(\mathbf{W}), \end{aligned}$$

Assuming independent priors for $\boldsymbol{\beta}$, $\boldsymbol{\phi}$, θ and $\mathbf{W}$, we have:

$$\begin{aligned} p(\mathbf{y}_o|\boldsymbol{\beta},\boldsymbol{\phi},\theta,\boldsymbol{\lambda},\mathbf{W},\mathbf{y}_f) &= \prod_{i=1}^{M}\prod_{j=1}^{m_i} c\left(F_1(y_{ij1}),...,F_K(y_{ijK})|\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\phi}\right) \\ &\times \prod_{k=1}^{K} p(y_{ijk}|\lambda_{i1k},\beta_{2k},\beta_{3k},\phi_k) \end{aligned}$$

and

$$\begin{aligned} p(\mathbf{y}_f|\boldsymbol{\beta},\boldsymbol{\phi},\theta,\boldsymbol{\lambda},\mathbf{W}) &= \prod_{i=1}^{M}\prod_{j=1}^{M_i-m_i} c\left(F_1(y_{ij1}),...,F_K(y_{ijK})|\boldsymbol{\lambda},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\phi}\right) \\ &\times \prod_{k=1}^{K} p(y_{ijk}|\lambda_{i1k},\beta_{2k},\beta_{3k},\phi_k). \end{aligned}$$

The posterior distribution of all unknown parameters has no close form and thus MCMC simulation might be applied. Assigning a Wishart prior to $\mathbf{W}$ and a Normal to the components of $\boldsymbol{\beta}$, provide full conditional with known forms for these parameters. Therefore, we can use Gibbs for sampling from them. The other parameters are sampled via Metropolis-Hastings algorithm. The full conditional $\mathbf{y}_f$ depends only on $p(\mathbf{y}_f|\boldsymbol{\beta},\boldsymbol{\phi},\theta,\boldsymbol{\lambda},\mathbf{W})$. Therefore, for simulating values from the distribution of $\mathbf{y}_f$, given the other parameters, is sufficient to simulate the pair $(\mathbf{y}_{ij1}^{(l)},\mathbf{y}_{ij2}^{(l)})$ from the Gaussian copula with Beta marginal distributions, for each l iteration of the algorithm, with $\boldsymbol{\beta}^{(l)}$, $\boldsymbol{\phi}^{(l)}$, $\theta^{(l)}$, $\boldsymbol{\lambda}^{(l)}$ and $\mathbf{W}^{(l)}$ for $j \notin s$, where s represents the selected sample.

The sampling process for the parameters of the model 1 is analogous. For simulating from $\mathbf{y}_{ijk}^{(l)}$, for $j \notin s$, we sampled from a Beta distribution with parameters μ_{ijk} and ϕ_k, where μ_{ijk} depends on λ_{i1k}, β_{2k} and β_{3k}.

3.2.2 Bayes Estimators

In this section are derived the Bayes estimators of the small area quantities of interest for the first application. The MCMC procedures provides a sample of size L of the predictive distribution of y_{ijk}, $j \notin s$. Therefore, it is possible to calculate point estimates (means or median) of any function of interest which involves y_{ijk}, $j \notin s$, as well as, a measure of variability of it, such as the posterior variances. Credibility intervals can also be provided.

The proportion of the correct answers for the k^{th} response variable in the i^{th} municipality can be written as:

$$\bar{Y}_{ik} = \frac{1}{\sum_{j=1}^{M_i} N_{ij}} \left(\sum_{j \in s} N_{ij} y_{ijk} + \sum_{j \notin s} N_{ij} y_{ijk} \right),$$

where N_{ij} is the number os students in the j^{th} school belong to the i^{th} municipality.

If the aim is to predict the mean $\bar{Y}_{ik}$, from the MCMC results, we can obtain L samples from the posterior distribution of $\bar{Y}_{ik}$: $l = 1, ..., L$:

$$\bar{Y}_{ik}^{(l)} = \frac{1}{\sum_{j=1}^{M_i} N_{ij}} \left(\sum_{j \in S} N_{ij} y_{ijk} + \sum_{j \notin S} N_{ij} y_{ijk}^{(l)} \right).$$

Thus the Bayes estimate of $\bar{Y}_{ik}$ under square loss is given by:

$$\hat{\bar{Y}}_{ik} = \frac{1}{L} \sum_{l=1}^{L} \bar{Y}_{ik}^{(l)},$$

$k = 1, 2$ and $i = 1, ..., M$.

To access the accuracy of the estimates provided by each model, the observed proportion of corrected answers for each subject $k = 1, 2$ was compared with the respective prediction of y_{ijk}, $j \notin s$ for both models.

3.2.3 Some Results

Models 1 and 2 were fitted, as well as the hierarchical beta regression with uncorrelated random effects. Table 1 contains a summary of the posterior distribution of the parameters for Model 2, with Gaussian copula fit. It should be noted that the posterior mean τ is quite high (0.629), indicating a high degree of association between the disciplines within the schools. The same is true regarding to the correlation of variables within the municipalities, represented by ρ_{12}: 0.693. These values show that the subjects should be jointly modeling. None of the credible intervals of coefficients regression contains zero, thus the auxiliary variables used are important for explanation of the responses and help to improve the predictions of the proportions of correct answers for both disciplines.

Table 2 contains the values of DIC, and its components for the two models and for each response variable for the univariate models. The individual models have better performance than the joint ones, because they have lower DIC and greater predictive likelihood. We compare only the estimates provided by the three models with the "true" proportions, because the schools were considered units and we have no direct estimates derived from the sampling design. The following quantities were used to compare the estimates provided for non-sampled schools:

- The absolute relative error (ARE), given by $ARE_{ijk} = |\hat{y}_{ijk} - p_{ijk}|/p_{ijk}$;
- The coefficient of variation, given by $CV_{ijk} = \sqrt{\hat{\sigma}_{ijk}}/\hat{y}_{ijk}$, where $\hat{\sigma}_{ijk}$ is the posterior variance of y_{ijk}.

Table 1: Summary of the posterior distribution of the parameters of Model 2.

Parameter	2.5%	50%	97.5%	Mean	Std.
β_{11}	-0.227	-0.147	-0.049	-0.144	0.046
β_{21}	-0.245	-0.159	-0.107	-0.164	0.035
β_{31}	0.112	0.185	0.253	0.183	0.036
β_{12}	-0.439	-0.327	-0.213	-0.327	0.060
β_{22}	-0.209	-0.127	-0.067	-0.130	0.037
β_{32}	0.104	0.177	0.245	0.176	0.036
ϕ_1	57.928	66.787	76.409	66.891	4.902
ϕ_2	54.109	62.307	71.302	62.569	4.495
σ_1^2	0.041	0.057	0.084	0.058	0.011
σ_{12}	0.039	0.062	0.099	0.064	0.016
σ_2^2	0.101	0.146	0.218	0.149	0.031
ρ_{12}	0.547	0.693	0.806	0.689	0.067
θ	0.798	0.835	0.865	0.834	0.017
τ	0.588	0.629	0.665	0.628	0.019

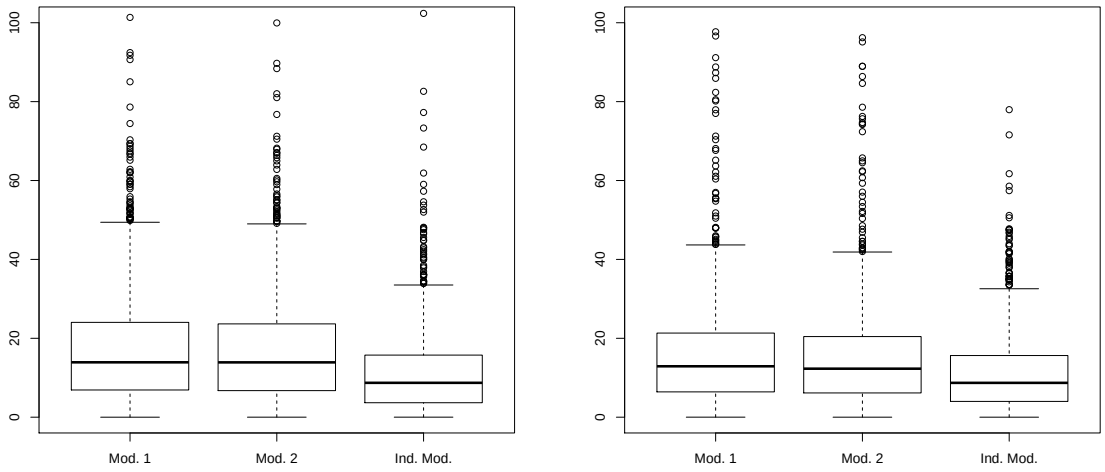
Table 2: DIC comparison criteria, penalized function associate with the number of parameters (p_D) and the logarithm of the predictive likelihood function ($\log p(\Psi)$) obtained by fitting the Hierarchical models without copula (Mod.1), with Gaussian copula (Mod.2) and the Univariate models (Mod. Univ.) for Portuguese and Mathematics tests

Model	DIC	p_D	$\log p(\Psi)$
Model 1	-2067.39	131.81	1099.60
Model 2	-1869.44	263.76	1066.60
Univ. Portuguese	-1115.99	48.54	582.27
Univ. Mathematics	-1049.28	66.12	557.70

The values of the quantities described above and presented in all charts bellow were multiplied by 100. Only those measures calculated for non-selected schools are presented in the graphs. Figure 1 compares the relative absolute errors. The results obtained from models 1 and 2 are quite similar. For the majority of the schools the ARE are less than 20 %. This can be considered as good performance of the models, since it can estimate the proportions with small relative errors.

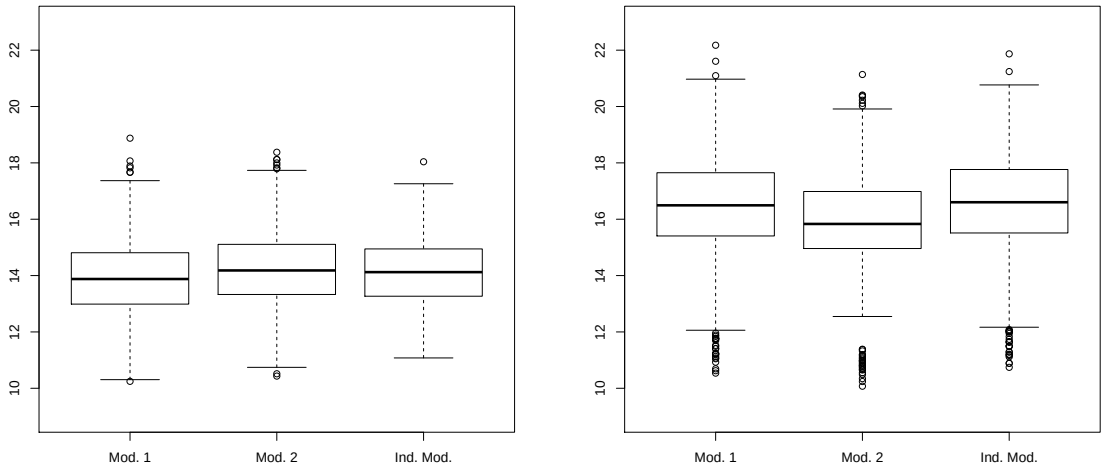
As can be seen in Figure 2, the coefficients of variation provided by the Model 2 are bit smaller for the proportions of correct answers of mathematics test than the others two models.

The Figures above show that by taking into account only the point estimates, the univariate models produce separate estimates close to what was observed and they could



(a) Portuguese (b) Mathematics

Figure 1: Box-plots of the ARE for (a) Portuguese test and (b) Mathematics using the Hierarchical models without copula (Mod.1), with Gaussian copula (Mod.2) and Univariate models (Univ. Mod.).



(a) Portuguese (b) Mathematics

Figure 2: Boxplots of the Coefficient of variation for (a) Portuguese and (b) Mathematics using the Hierarchical models without copula (Mod.1), with Gaussian copula (Mod.2) and Univariate models (Univ. Mod.).

be preferred because they are easier to fit than the others. However, the coefficients of variation produced by Model 2 are smaller than those obtained by other models for the proportion of correct answers in mathematics and close to what was found by the others in the Portuguese, i.e, as far as variability is concerned, the hierarchical model with copula would be a good adjustment option. The credibility intervals of all fitted models contain about 95% of the observations.

3.3 Application 2

The design sampling considered in this second application are more complex than the one considered in the first application. In each municipality of Rio de Janeiro State was selected a sample of schools with equal probability, and within schools was selected a sample of students. We suppose that only the sampled students do the Portuguese and Mathematics tests.

In this exercise, the average proportions of correct answers in both disciplines for each selected school are direct estimates, based on a small sample of students. The main aim is to estimate these proportions for the non-sampled schools and to reduce the errors for the sampled schools. Thus, the school is considered the small areas and it is applied a multivariate hierarchical beta area model, containing two parts: one relates the direct estimates with parameters of area, the other relates these parameters to the auxiliary variables.

Unlike the first application, the response variables have sample error that may be related to the area sample size. To consider this feature, a modification in the multivariate hierarchical model is proposed in the equation of the observations. Because it is natural to think that the variance of the estimate increases when sample size decreases, it is proposed the following two-level model:

$$y_{ijk} \sim beta(\mu_{ijk}, \phi_{ijk}),$$

where y_{ijk} is the direct estimate (based on the sampling design) of the expected proportion of the correct answers of the discipline k^{th}, of the j^{th} school in the i^{th} municipality, for $j = 1, ..., m_i,\ i = 1, ..., M,\ k = 1, ..., K$. Thus, the parameter of dispersion ϕ_{ijk} assumes different value for each sampled school, and its value depends on the sample size through the following function:

$$\phi_{ijk} = \gamma_k(n_{ij} - 1),$$

where n_{ij} is the size of the j^{th} school for the i^{th} municipality.

For the condition $\phi_{ijk} > 0$ be satisfied, we must have $n_{ij} \geq 2$ and $\gamma_k > 0$. The common factor γ_k ensures that if two schools have the same proportion of correct answers and equal sampling fraction, their variances will be different and that with smaller sample size will have higher variance. Moreover, when y_{ijk} is a proportion and $\gamma_k = 1$, it follows that $Var(y_{ijk}) = \frac{\mu_{ijk}(1-\mu_{ijk})}{n_{ij}}$, which is the variance of the proportion under simple random sampling.

Thus, the following model can be considered for the selected schools:

$$\begin{aligned} y_{ijk} &\sim Beta(\mu_{ijk}, \phi_{ijk}), \quad j = 1, ..., m_i, \quad i = 1, ..., M \\ g(\mu_{ijk}) &= \lambda_{i1k} + x_{ij2}\beta_{2k} + x_{ij3}\beta_{3k}, \quad k = 1, ..., K \\ \lambda_{i1k} &= \beta_{1k} + \nu_{i1k}, \quad \nu_{i1k} \sim N(0, \sigma_k^2), \end{aligned} \tag{6}$$

where only the intercepts are assumed to be random, m_i is the number of selected schools for the i^{th} municipality, $\phi_{ijk} = \gamma_k(n_{ij} - 1)$, and n_{ij} is the sample size of the students for the j^{th} school in the i^{th} municipality.

As the information of all students and schools is available on the Brazilian micro-data test, it is possible to calculate the true observed proportions of the selected schools and compare them with the direct estimates and those provided by the hierarchical model.

This proposed model can only be applied to the selected schools because we must have information on the sample size. In the following section is presented the inference process on the parameters of the model (6) and the indirect estimators of the sampled and not sampled areas.

3.3.1 Inference

Let $\mathbf{W} = \mathbf{\Sigma}^{-1}$. The posterior density of all model parameters, assuming independent priors for the parameters of the model (6), can be written as:

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{W}|\mathbf{y}) &\propto p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{W}) p(\boldsymbol{\lambda}|\boldsymbol{\beta}, \mathbf{W}, \boldsymbol{\gamma}) p(\boldsymbol{\beta}) p(\boldsymbol{\gamma}) p(\mathbf{W}) \\ &\propto p(\boldsymbol{\beta}) p(\mathbf{W}) p(\mathbf{y}|\boldsymbol{\lambda}, \boldsymbol{\gamma}) p(\boldsymbol{\lambda}|\boldsymbol{\beta}, \mathbf{W}) \times \left\{ \prod_{k=1}^{K} p(\boldsymbol{\gamma}_k) \right\} \end{aligned}$$

where

$$p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{W}) = p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) \propto \prod_{i=1}^{M} \prod_{j=1}^{m_i} \prod_{k=1}^{K} p(y_{ijk}|\lambda_{i1k}, \beta_{2k}, \beta_{3k}, \gamma_k),$$

and

$$p(\boldsymbol{\lambda}|\boldsymbol{\beta}, \mathbf{W}, \boldsymbol{\gamma}) = p(\boldsymbol{\lambda}|\boldsymbol{\beta}, \mathbf{W}) = \prod_{i=1}^{M} p(\boldsymbol{\lambda}_{i1}|\boldsymbol{\beta}_1, \mathbf{W})$$

$$\propto \prod_{i=1}^{M} |\mathbf{W}|^{1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\lambda}_{i1} - \boldsymbol{\beta}_1)^T \mathbf{W} (\boldsymbol{\lambda}_{i1} - \boldsymbol{\beta}_1)\right\},$$

Analogously to what was discussed in the first application, the posterior distribution above has no close form. As in the hierarchical model, the full conditional of $\mathbf{W}$ and $\boldsymbol{\beta}_1$ have close forms with the assumption that these parameters respectively follow the bivariate normal and Wishart distributions. The process of obtaining the full conditional will be omitted because it is analogous to the ones previously presented.

To sample from the parameters $\mathbf{W}$ and $\boldsymbol{\beta}_1$ was used the Gibbs sampler, while the others was employed the Metropolis-Hastings algorithm.

The more general model which makes use of copulas, had convergence problems and because of that its results are not shown.

3.3.2 Small area Estimators

The process of modeling and inference presented below is with only respect to the sampled schools, for which the indirect estimates provided by the model are given by $y_{ijk}^{(l)}$, $j \in s$, $k = 1, 2$, $i = 1, ..., M$. This is obtained by jointly simulating the pairs $(y_{ij1}^{(l)}, y_{ij2}^{(l)})$ from the Beta distributions $(\mu_{ijk}^{(l)}, \phi_{ijk}^{(l)})$, where $\mu_{ijk}^{(l)} = g^{-1}\left(\lambda_{i1k}^{(l)} + \beta_{2k}^{(l)} x_{ij2} + \beta_{3k}^{(l)} x_{ij3}\right)$ and $\phi_{ijk}^{(l)} = \gamma_k^{(l)}(n_{ij} - 1)$ for $k = 1, 2$, $i = 1, ..., M$ and $j = 1, ..., m_i$. The quantity $\mu_{ijk}^{(l)}$ can be also used as estimator. The choice between $\mu_{ijk}^{(l)}$ and $y_{ijk}^{(l)}$ depends on the researcher's interest: if we want to estimate what would be predicted by the survey, we should use $y_{ijk}^{(l)}$, if you want to know how much, on average, the students of the j^{th} school scores in each discipline, we should use $\mu_{ijk}^{(l)}$.

No model was assumed for the non-selected schools, nevertheless it is also necessary define the estimators for these schools. If there is information on the auxiliary variables for these schools, the estimate of expected proportion in each non-selected school at each (l) sample point of the posterior distribution is given by:

$$\mu_{ijk}^{(l)} = g^{-1}\left(\lambda_{i1k}^{(l)} + \beta_{2k}^{(l)} x_{ij2} + \beta_{3k}^{(l)} x_{ij3}\right).$$

Since we have L sample points from the posterior distribution of μ_{ijk}, we can obtain the point estimates and the credibility intervals for μ_{ijk}; $j \notin s$. Therefore, L samples from the posterior distribution of μ_{ik} for the k^{th} discipline in the i^{th} municipality is given by:

$$\mu_{ik}^{(l)} = \frac{1}{\sum_{j=1}^{M_i} N_{ij}} \left(\sum_{j \in S} N_{ij} \mu_{ijk}^{(l)} + \sum_{j \notin S} N_{ij} \mu_{ijk}^{(l)}\right),\ l = 1, ..., L.$$

Thus, the Bayes estimators of μ_{ijk} under square loss is given by:

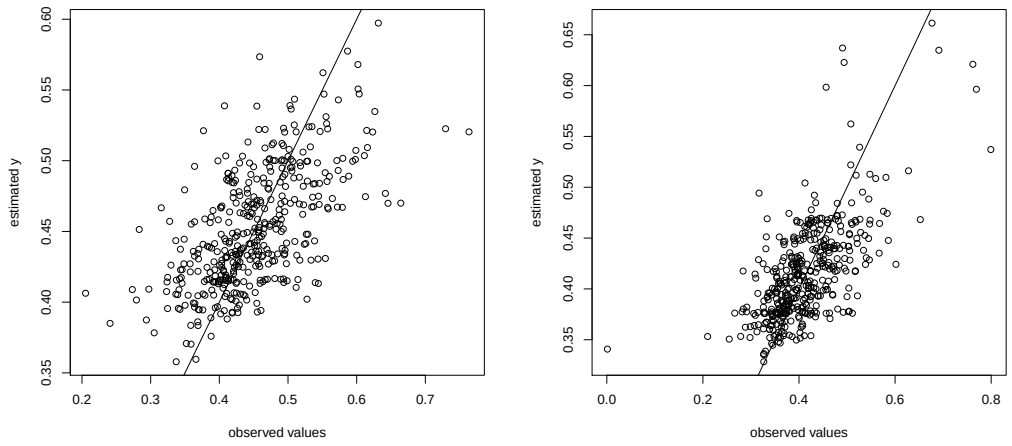
$$\hat{\mu_{ik}} = \frac{1}{L}\sum_{l=1}^{L}\mu_{ik}^{(l)}.$$

3.3.3 Some Results

The main aims of modeling the proportions of correct answers are to reduce variability of direct estimates derived from the sampling design and to obtain estimates for non-sampled schools with good accuracy. The direct estimators can be only obtained for selected schools. The multivariate model is able to provide estimates for all schools, but we need to check its model adequacy.

The 95% credible intervals of the predictive proportions by the replica $y_{ijk}^{(l)}$, $j \in S$, respectively contains 98.1% and 95.7% of the observed values for Portuguese and Mathematics disciplines.

We can also compare the observed value with $y_{ijk}^{(l)}$, $j \in S$ to access the adequacy of the model, as presented in Figure 3.



(a) Portuguese, sampled schools

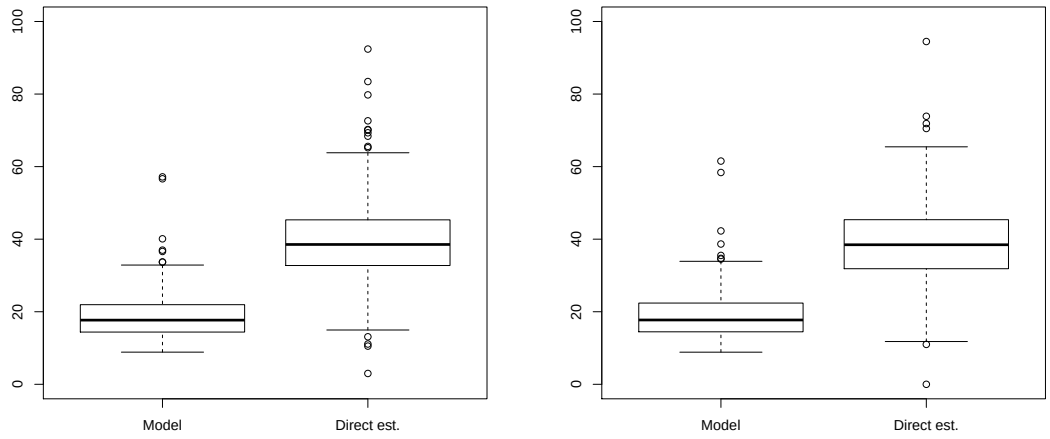
(b) Mathematics, sampled schools

Figure 3: Plot of the proportions of corrected answers against the posterior means of y_{ijk}: (a) Portuguese; (b) Mathematics

Analyzing the figures above, we can conclude that the multivariate hierarchical beta model produced reasonable estimates for the sampled schools.

The reduction of the variability of direct estimates by application of the model can be verified by assessing the coefficients of variation (CV) provided by the direct estimator and by the estimator obtained by employing the model. Figure 4 summarizes the distribution of the coefficients of variation obtained from the two estimators. Clearly the CV's

generated by model are much lower than those obtained by the direct estimation.



(a) Portuguese (b) Mathematics

Figure 4: Coefficients of variation of the direct estimator and the posterior mean of the quantity of interest for the sampled schools (a) Portuguese and (b) Matemathics.

4 Concluding Remarks and Suggestions for Future Work

The models proposed have the advantage of keeping the response variables in their original scale. Another advantage refers to the use of copulas which are marginal-free, i.e, the degree of association of variables is preserved whatever the marginal distributions are. Thus, if two indexes are correlated whatever the marginal adopted, the measure of dependence is the same. The use of copula functions in beta marginal regressions allows to jointly analyze the response variables, by taking advantage of their dependency structure and keeping the variables in their original scale. The application of multivariate models with Beta responses is an appealing alternative to models that require transforming the original variables. The choice between the proposed models and its competitors in the literature should be guided by the goals of the researcher, who must observe the predictive power and the goodness of fit of them. The disadvantage of models that uses copulas is their time consuming for simulating samples from the posterior distributions of the model parameters or functions of them.

In Section 2, we propose a multivariate hierarchical model with two levels where the variables are correlated in the first level with the aid of a copula function. Despite being applicable in general situations, this model has been developed especially for the small

area estimation problem to allow exchange of information between the areas or small domains of interest. It is assumed that the random effects of the same area are correlated and the random effects of different areas has the same variance-covariance matrix.

In the first application, using the DIC and predictive likelihood criteria, the multivariate model performs worse than the separate beta models. The Gaussian copula model tends to overestimate the proportions of interest. This model must be investigated and it is worth note that only the Gaussian copula was used and this may not be adequate for these data. In Application 2, the multivariate hierarchical model was able to estimate the expected proportions for non-sampled schools and also presented a significant reduction of the coefficients of variance when compare to the direct estimates.

Sample household surveys are important sources of potential applications of the models proposed in this work. Examples of variables measured in the range $(0, 1)$ are the occupancy rate and the poverty gap, which is a ratio between the total incomes of individuals below the line of poverty and the sum of all incomes of the population. These variables are important measures for planning and knowledge of the population conditions, but are rarely available for small geographic levels or population subgroups for intercensus periods. The prediction of these poverty index could be done by employing the models proposed in this paper.

It is important to note that this work focuses on building multivariate regression models in which the marginal distributions are Beta. It points out its advantages over corresponding univariate models and the difficulties of estimating their parameters. However, the theory of copula functions can be applied to any multivariate models that can be built for any known marginal distributions, allowing that the distributions of response variables involved be different. We can even have continuous and discrete variables in the same model. To build a model for others distributions is straightforward, but each model has a peculiar and practical feature, and the estimation process should always be taken into account when we propose a new model. In the specific case of the Beta model, has been adopted the mean and the dispersion as the model parameters, where the latter parameter controls the variance. Other parameterizations are possible, but could lead to additional difficulties. Various strategies can be defined by the researcher, according to the available database, some important ones are: first fixe the marginal and then obtain the more appropriate copulas; estimate models with different copulas and marginal and decide what is "the best" model by applying a model comparison approach.

Another worth point to be mentioned is that in practical situations where the response variables can have zeros or ones values, the Beta distribution will not be adequate. One possible way of circumventing this problem is to use a mixture of distributions, so that the zeros and ones can be accommodated. Ospina and Ferrari (2010) proposes inflated beta regression models to fit data with such feature. We have not considered omission in the explanatory variables in our model formulation, which could be another possible extension of the models proposed here.

References

Datta, G. S., Day, B., Basawa, I., 1999. Empirical best linear unbiased and empirical bayes prediction in multivariate small area estimation. Journal of Statistical Planning and Inference 75, 169–179.

Fabrizi, E., Ferrante, M. R., Pacei, S., Trivisano, C., 2011. Hierarchical bayes multivariate estimation of poverty rates based on increasing thresholds for small domains. Computational Statistics and Data Analysis 4 (1), 1736–1747.

Fay, R. E., 1987. Application of multivariate regression to small domain estimation. In: Platek, R., Rao, J., Srndal, C., Singh, M. (Eds.), Small Area Statistics. Wiley, New York, pp. 91–102.

Ferrari, S. L. P., Cribari-Neto, F., 2004. Beta regression for modelling rates and proportions. Journal of Applied Statistics 31 (7), 799–815.

Jiang, J., 2007. Linear and Generalized Linear Mixed Models and Their Applications. Springer Series in Statistics. Springer, New York.

Melo, T. F. N., Vasconcellos, K. L. P., Lemonte, A. J., 2009. Some restriction tests in a new class of regression models for proportions. Computational Statistics and Data Analysis 53, 3972–3979.

Nelsen, R. B., 2006. An Introduction to Copulas, 2nd Edition. Springer, New York.

Olkin, I., Liu, R., 2003. A bivariate beta distribution. Statistics and Probability Letters 62, 407–412.

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

Pfeffermann, D., Moura, F., Silva, P., 2006. Multi-level modeling under informative sampling. Biometrika 93, 943–959.

Rao, J. N. K., 2003. Small area estimation. Wiley, New York.