BIVARIATE BETA REGRESION MODELS:
a Bayesian approach applied to educational data

Edilberto Cepeda-Cuervo.[1]

Departamento de Estadística

Universidad Nacional de Colombia, Bogotá, Colombia

Jorge Alberto Achcar[2]

Departamento de Medicina Social

Universidade de São Paulo, FMRP, Brazil

Liliana Garrido Lopera.[3]

Departamento de Matemáticas

Universidad de los Andes, Bogotá, Colombia

[1] email: ecepedac@unal.edu.co

[2] email: achcar@fmrp.usp.br

[3] email: bgarrido@uniandes.edu.co

Summary

In this paper we propose a bivariate beta regression model, defining the beta distribution derived from Farlie-Gumbel-Morgenstern (FGM) copulas. This model could be a good alternative to analyze pairs of proportions, when they are not independent. To fit the proposed models we apply standard existing MCMC (Markov Chain Monte Carlo) methods to simulate samples for the joint posterior of interest, using the Bayesian methodology proposed by Cepeda and Gamerman (2001) and Cepeda and Gamerman (2005). Two examples are introduced to illustrate the proposed methodology: an example with simulated bivariate data and an example with a real data set.

Key words*: Beta distribution; Beta regression models; bivariate random variables; MCMC methods; Bayesian methodology.*

1. Introduction

In many applications in different areas as medicine, education, economics, ecology among many others, we could have a response given by a rate or proportion that is limited to a value in the interval $(0, 1)$. Usually, we also have the presence of a vector of covariates associated to each unit.

A flexible and natural candidate distribution to model this kind of data is given by the beta distribution (see for example, Johnson, Kotz and Balakrishnan, 1998, p. 235).

As an application and motivation for this paper, we consider a study to evaluate the quality of education in different regions, where in the evaluation of the schools performance in mathematics, language, natural sciences among many other school areas, a number between 0 and 5 (or any other positive integer) is assigned as a measure of the student performance. In this case, the measure assigned to each student could be expressed as a number between zero and one and a beta distribution could be used to analyze the data in the presence or not of a vector of covariates. Some applications of the beta distribution are presented and analyzed by Bury (1999).

Many studies to determine the quality of the educational systems in different countries were developed in different contexts. The Program for International Student Assessment (PISA) was designed and launched by OECD (Organization for Economic Co-operation and Development) by the end of the nineties decade as an international and comparative index to evaluate the performance of the school children that would enable countries to improve their educational systems.

In 2009 the PISA program was applied to 57 countries, including 37

OECD countries and other 27 countries called "partner countries". In this case, the explanatory variables measured the ways of learning as well as the influence of social and family environmental factors, school material, teaching staff, expectation and inclinations of the teenagers to learn, among others topics (Sancho, 2006; Cepeda, 2005; Donoso, 2002).

Similar studies where the student's performance or the educational system is associated with a beta distribution with explanatory variables like family environments, socioeconomic or scholar variables, have been developed in countries as Colombia, Chile, Spain and the United States of America.

In these models, if Y is the variable of interest with beta distribution, we relate a vector of covariates $X' = (X_1, \ldots, X_k)$ by the link function $g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$, with $\mu_i = E(Y_i)$, $i = 1, 2, \ldots, n$ where n is the sample size, g is an appropriate real valued function, $\boldsymbol{\beta'} = (\beta_1, ..., \beta_k)$ is a vector of regression parameters and $\mathbf{x}_i'$ is the vector of covariates associated to the i-th individual.

This modeling approach was introduced by Cribari-Neto (2005), although more general models were proposed by Cepeda (2001) considering joint modeling of the mean and variance or dispersion parameters in the biparametric exponential family, including joint modeling of the mean and dispersion parameters in the beta distribution.

Inferences for the beta regression models have been discussed by many authors under a classical inference approach (see for example, Paolino, 2001, or Ferrari and Cribari-Neto, 2004) or under Bayesian approach (see for example, Cepeda, 2001; Branscum et al., 2007).

In some situations, we could have two proportion responses associated to the same unit, as it is the case of medical longitudinal data where a patient could have a proportion of virus in the blood measured before and after receiving a treatment. In this case we can not assume independence between the responses and we need a bivariate beta distribution to analyze the data (see for example, Olkin and Liu, 2003).

The dependence between the observed proportions could also be studied considering the use of copula functions (see for example, Nelsen, 1999).

In this paper we present a Bayesian analysis for univariate and bivariate regression models. In the case of bivariate regression models, we assume a special copula function: the Farlie-Gumbel-Morgenstern (FGM) copula which is appropriate to fit data with weak dependences. We get the posterior summaries of interest using standard MCMC (Markov Chain Monte Carlo) methods to simulate samples for the joint posterior distribution of interest (see for example, Gelfand and Smith, 1990).

The paper is organized as follows: In section 2.1, general concepts on the beta distribution are introduced; in section 2.2, we present some concepts on the copula functions; in section 2.2.1, we present a bivariate beta distribution derived from a FGM copula; in section 3.1, we introduce a general joint mean and dispersion (variance) beta regression model; in section 3.2, we introduce a bivariate beta regression model; in section 4, we introduce a Bayesian approach for the model; in section 5, we present a simulation study; in section 6, we present a real data set; finally in section 7, we present some conclusions.

2. The beta distribution

2.1. Univariate beta distribution

A random variable Y has a beta distribution if its density function is given by

$$f(y|p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} I_{(0,1)}(y), \tag{1}$$

where $p > 0$, $q > 0$ and $\Gamma(.)$ denotes the gamma function. If Y is a random variable with beta distribution, the mean $\mu = E(Y)$ and variance $\sigma^2 = Var(Y)$ are given respectively by

$$\mu = \frac{p}{p+q}, \tag{2}$$

$$\sigma^2 = \frac{p\,q}{(p+q)^2(p+q+1)}. \tag{3}$$

Many observations could be assumed to have a beta distribution. For example, the income inequality or the land distribution when they are measured using the Gini index proposed by Atkinson(1970). Another quantity that can be analyzed using a beta distribution is the performance of a student in areas as mathematics , natural sciences or literature. In this case, if the performance X takes values in an interval (a, b), the random variable $Y = (X-a)/(b-a)$ can be assumed to have a beta distribution. In this case, there are household socioeconomic variables that have fundamental impact on the cognitive achievement of the students. For example, the level of student achievement is closely related to educational levels of their parents and the number of hours devoted to study a subject. Thus, the beta regression model can be appropriate to explain the behavior of the school performance as a function of many associate factors. Some reparametrizations of the beta distribution given in (1) can be more appropriate. As a first one, let us consider $\phi = p + q$; we can see that $p = \mu\phi$, $q = \phi(1 - \mu)$ and $\sigma^2 = \frac{\mu(1-\mu)}{\phi+1}$. In this case, ϕ can be interpreted as a precision parameter in the sense that, for fixed values of μ, larger values of ϕ correspond to smaller values of the variance of Y. This reparametrization given in Ferrari and Cribari-Neto (2004), had already been early introduced in the literature, for example in Jorgensen (1997) or in Cepeda (2001). With this reparametrization, the density of the beta distribution (1) can be rewritten as

$$f(y|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} I_{(0,1)}(y). \tag{4}$$

In this case, the mean and dispersion parameters can be modeled as function of explanatory variables as it was proposed by Cepeda(2001).

The beta distribution given in (1) can also be reparametrized as a function of the mean and variance, in the following way:

$$\begin{aligned} p &= \frac{(1-\mu)\mu^2 - \mu\sigma^2}{\sigma^2} && (5) \\ q &= \frac{(1-\mu)[\mu - \mu^2 - \sigma^2]}{\sigma^2} && (6) \end{aligned}$$

Although considering (1) as a function of μ and σ^2 can result in a complex expression, joint modeling of the mean and variance can be easily obtained applying the Bayesian methodology proposed by Cepeda(2001). Sometimes the joint modeling of the mean and variance can be more appropriate than joint modeling the mean and the dispersion parameter given that the parameters of the regression models will be more easily interpreted.

2.2. Bivariate beta distribution

To capture the existing dependence between two observed proportions, we build a bivariate beta distributions, using copula functions. A popular copula function considered for the study of the dependence structure of two random variables is based on the Farlie-Gumbel-Morgentern (FGM) copula given by

$$C(u_1, u_2; \theta) = u_1 u_2 (1 + \theta(1 - u_1)(1 - u_2)), \tag{7}$$

where $u_1 = F_{(v_1)}$ and $u_2 = F_{(v_2)}$ are the marginal distribution functions and θ, $-1 < \theta < 1$, is a measure of the dependence between them.

For interpretation of the dependence parameter θ, can be used the relationship between it and the association coefficients Kendall's Tau (τ) and Spearman's Rho (ρ), given by the equations:

$$\begin{aligned} \tau &= 4 \int\int C(u_1, u_2) dC(u_1, u_2) - 1 \\ &= 4\Big(\frac{\theta}{18} + \frac{1}{4}\Big) - 3 = \frac{2\theta}{9}, \end{aligned} \tag{8}$$

and

$$\begin{aligned} \rho &= 12 \int\int u_1 u_2 dC(u_1, u_2) - 3 \\ &= 12\Big(\frac{1}{4} + \frac{\theta}{36}\Big) - 1 = \frac{\theta}{3} \end{aligned} \tag{9}$$

(see for example, Nelsen, 1999).

Other copula functions could be considered to analyze bivariate continuous data, see Gumbel copula (Gumbel, 1960) and Clayton copula (Clayton, 1978), among others.

2.2.1. A bivariate beta distribution

In this section, a bivariate beta distribution is derived from the FGM copula function (7), assuming that $u_1 = F_1(y_1)$ and $u_2 = F_2(y_2)$ are the marginal beta distributions, given by

$$F_k(y_k) = P(Y_k \leq y_k) = \int_0^{y_k} f_k(t; \mu_k, \phi_k) dt \quad k = 1, 2. \tag{10}$$

where f_k, $k = 1, 2$ are beta density functions. Thus, the bivariate beta distribution and density functions are given by

$$F_I(y_1, y_2) = F_1(y_1)F_2(y_2)[1 + \theta[1 - F_1(y_1)][1 - F_2(y_2)]], \tag{11}$$

and

$$\begin{aligned} f_I(y_1, y_2) &= \frac{\partial F_I(y_1, y_2)}{\partial y_1 \partial y_2} \\ &= f_1(y_1)f_2(y_2) + \theta f_1(y_1)f_2(y_2)[1 - 2F_1(y_1)][1 - 2F_2(y_2)] \quad (12) \end{aligned}$$

respectively.

The bivariate beta distribution function has five parameters: μ_1 and μ_2 for the means, ϕ_1 and ϕ_2 for the precisions, and θ for the dependence parameter. If $\theta = 0$ the bivariate beta density function is given by $f_I(y_1, y_2) = f_1(y_1)f_2(y_2)$, which shows that the random variables Y_1 and Y_2 are independent.

3. Beta regression models

3.1. Univariate Beta regression models

The study of the beta regression models has received great attention following the work of Cepeda(2001), where joint mean and dispersion beta regression models are proposed, under a Bayesian approach. In these models, it is assumed that the interest variable $\mathbf{Y}_i$, $i = 1, 2, \ldots, n$, has a beta distribution, where the mean and dispersion models are given by (13) and (14), respectively, in which $\boldsymbol{\beta}^{'} = (\beta_0, \beta_1, ..., \beta_k)$ and $\boldsymbol{\gamma}^{'} = (\gamma_0, \gamma_1, \ldots, \gamma_p)$ are the parameter vectors, and $\mathbf{x}_i$ and $\mathbf{z}_i$ the mean and dispersion explanatory variables.

$$\begin{aligned} \text{logit}(\mu_i) &= \mathbf{x}_i^{'}\boldsymbol{\beta}, & (13) \\ \log(\phi_i) &= \mathbf{z}_i^{'}\boldsymbol{\gamma}, & (14) \end{aligned}$$

Assuming the same reparametrization of the beta distribution, $\mu = p/(p+q)$ and $\phi = p+q$, Ferrari and Cribari-Neto (2004) proposed a particular case of these models assuming constant dispersion parameter. In more recent paper, joint mean an dispersion beta regression models were introduced by Smithson and Verkuilen (2006) and Simas et al. (2010), under a classical approach. At the same time, nonlinear beta regression models were proposed by Cepeda and Achcar (2010), assuming that the mean model is given by

$$\mu_i = \tfrac{\beta_0}{1+\beta_1 \exp(\beta_2 x_i)} \tag{15}$$

and the dispersion model by (14), in the context of double generalized nonlinear models. This model was applied in the analysis of the schooling rate in Colombia, for the period ranging from 1991 to 2003. Nonlinear regression models were also considered by Simas et al. (2010).

3.2. Bivariate beta regression

The joint mean and dispersion bivariate beta regression models are defined assuming that the mean and dispersion parameter, μ_k and ϕ_k, $k = 1, 2$, can be modeled as functions of the explanatory variables. Thus, we assume that the random variables (Y_{1i}, Y_{2i}), $i = 1, 2, \ldots, n$, have a bivariate beta distribution, with mean and dispersion models given by

$$\begin{aligned} h_k(\mu_{ki}) &= \mathbf{x}_{ki}^{'}\boldsymbol{\beta}_k & (16)\\ g_k(\phi_{ki}) &= \mathbf{z}_{ki}^{'}\boldsymbol{\gamma}_k, \quad k = 1, 2 & (17)\end{aligned}$$

where h_k and g_k are appropriate real valued functions, $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kr_k})$ and $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kr_k})$. Thus, the likelihood function is given by

$$\begin{aligned} L(\theta, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) &= \prod_{i=1}^{n} f_I(y_{1i}, y_{2i} \mid x_i) & (18)\\ &= \prod_{i=1}^{n} f_1(y_{1i}) f_2(y_{2i})[1 + \theta[1 - 2F_1(y_{1i})][1 - 2F_2(y_{2i})]] \end{aligned}$$

where

$$f_k(y_{ki}) = \frac{\Gamma(\phi_k) y_{ki}^{\mu_{ki}\phi_k - 1}(1 - y_{ki})^{(1-\mu_{ki})\phi_k - 1}}{\Gamma(\mu_{ki}\phi_k)\Gamma[(1 - \mu_{ki})\phi_k]}, \tag{19}$$

and

$$F_k(y_{ki}) = \int_0^{y_{ki}} \frac{\Gamma(\phi_k) t^{\mu_{ki}\phi_k - 1}(1 - t)^{(1-\mu_{ki})\phi_k - 1} dt}{\Gamma(\mu_{ki}\phi_k)\Gamma[(1 - \mu_{ki})\phi_k]}. \tag{20}$$

with μ_{ki} and ϕ_{ki} given by (16) and (17). If Y_1 and Y_2 are independent, $\theta = 0$ and $L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) = \prod_{i=1}^{n} f_1(y_{1i}) f_2(y_{2i})$.

4. Bayesian methodology

4.1. Univariate regression models

In this section we apply the Bayesian methodology and the MCMC algorithm proposed in Cepeda (2001) and Cepeda et al. (2001, 2005), in the framework of double generalized regression model, to fit a bivariate beta regression model. As in these works, to implement a Bayesian approach to estimate the parameters of the joint beta regression model, we need to specify a prior distribution for the parameters. Thus, if $L(\boldsymbol{\Theta})$ denotes the likelihood function and $p(\boldsymbol{\Theta})$ the joint prior distribution, where $\boldsymbol{\Theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$, the posterior distribution is given by $\pi(\boldsymbol{\Theta}|\text{ data}) \propto L(\boldsymbol{\Theta})p(\boldsymbol{\Theta})$. However, given that assuming normal prior distributions, the posterior distribution $\pi(\boldsymbol{\Theta}|\text{ data})$ is analytically intractable and it is not easy to generate samples from it, Cepeda(2001) proposed to get samples of $\boldsymbol{\Theta}$ using an iterate alternating algorithm, sampling $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ from the posterior conditional distributions $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma},\text{ data})$ and $\pi(\boldsymbol{\gamma}|\boldsymbol{\beta},\text{ data})$, for which it is necessary to build normal transition kernels q_1 and q_2, given that these conditional distributions are also analytically intractable. We assume normal prior distributions for the regression parameters, given that their maximum likelihood estimators are asymptotically normal and given that for large values of the variance of the prior distribution there are no remarkable changes in the posterior distribution.

To build the kernel transition functions we need to define working observation variables to approximate $h(\mu_i)$ and $g(\phi_i)$ around the current values of μ and ϕ, respectively. This variables are defined as first order Taylor approximations of the real functions $h(t_1)$ and $g(t_2)$, where t_1 and t_2 are random variables such that $E(t_1) = \mu$ and $E(t_2) = \phi$. Thus, given that $E(t_1) = \mu$ for $t_1 = Y$, if the mean model is given by (13), the working observational variable is defined by

$$\tilde{y}_i = x_i^{'}\boldsymbol{\beta}^{(c)} + \frac{y_i - \mu_i^{(c)}}{(\mu_i^{(c)})(1-\mu_i^{(c)})}, \ i = 1, 2, ..., n, \tag{21}$$

where $\mu^{(c)}$ and $\boldsymbol{\beta}^{(c)}$ are the current values of μ and $\boldsymbol{\beta}$. Thus, assuming that (21) has a normal distribution and assuming conditional normal prior distribution $\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N(b, B)$, the kernel transition function q_1 is given by the posterior distribution obtained from the combination of the prior distribution with the working observation model $\tilde{y}_i \sim N(x_i^{'}\boldsymbol{\beta}, \tilde{\sigma}_i^2)$, where $\tilde{\sigma}_i^2 = \text{Var}(\tilde{y}_i)$. That is, by

$$q_1(\boldsymbol{\beta}|\boldsymbol{\beta}^{(c)}, \boldsymbol{\gamma}^{(c)}) = N(\mathbf{b}^*\ , \mathbf{B}^*), \tag{22}$$

where

$$\begin{aligned} \mathbf{b}^* &= \mathbf{B}^*(\mathbf{B}^{-1}\mathbf{b} + \mathbf{X}^{'}\boldsymbol{\Sigma}^{-1}\tilde{Y}) \\ \boldsymbol{\beta}^* &= (\mathbf{B}^{-1} + \mathbf{X}^{'}\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1} \end{aligned}$$

and where $\boldsymbol{\Sigma}$ is a diagonal matrix with diagonal entries $\tilde{\sigma}_i^2$, $i = 1, 2, ..., n$, (see Cepeda and Gamerman, 2001, 2005). Thus, the values of $\boldsymbol{\beta}$ from the posterior distribution sample of $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma})$ are proposed from the transition kernel defined in equation (22).

As the full conditional distribution $\pi(\boldsymbol{\gamma}|\boldsymbol{\beta})$ is analytically intractable and it is not easy to generate samples from it, we need to build a kernel transition q_2 to propose the values of $\boldsymbol{\gamma}$ from the posterior distribution of $\boldsymbol{\Theta}$. Given that $E(\boldsymbol{t_i}) = \phi_i$ for $\boldsymbol{t_i} = \frac{(p_i+q_i)}{p_i}\boldsymbol{Y}_i$, if the precision model is given by (14), the working observational variable (23) is obtained from the first order Taylor approximation around of current value of ϕ_i, given by the current values of the precision regression models $\boldsymbol{\gamma}^{(c)}$, given by

$$\tilde{y}_i = \boldsymbol{z}_i^{'}\boldsymbol{\gamma}^{(c)} + \frac{y_i}{\mu_i} - 1, \;\; i = 1, 2, ..., n. \tag{23}$$

Thus, assuming that the observational working variable (23) has a normal distribution and given that the conditional prior distribution is given by $\boldsymbol{\gamma}|\boldsymbol{\beta} \sim N(\mathbf{g}, \mathbf{G})$, the normal transition kernel q_2 is given by the posterior distribution obtained from the combination of the prior distribution with the working observational model $\tilde{y}_i \sim N(z_i^{'}\boldsymbol{\gamma}, \tilde{\sigma}^2)$, where $\tilde{\sigma}_i^2 = \text{Var}(\tilde{y}_i)$. That is,

$$q_2(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(c)}, \boldsymbol{\beta}^{(c)}) = N(\mathbf{g}^* \;, \mathbf{G}^*), \tag{24}$$

where

$$\begin{aligned} \mathbf{g}^* &= \mathbf{G}^*(\mathbf{G}^{-1}\mathbf{g} + \mathbf{Z}^{'}\Psi^{-1}\tilde{Y}), \\ \mathbf{G}^* &= (\mathbf{G}^{-1} + \mathbf{Z}^{'}\Psi^{-1}Z)^{-1}. \end{aligned}$$

and Ψ is a diagonal matrix with entries $\tilde{\sigma}_i^2$ for $i = 1, 2, ..., n$. Samples of $\boldsymbol{\gamma}$ from the posterior distribution $\pi(\boldsymbol{\beta}, \boldsymbol{\gamma})$, are obtained from the transition kernel function q_2.

With the transition kernels given by (22) and (24), the components $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ of $(\boldsymbol{\beta}, \boldsymbol{\gamma})'$ are updated as follows:

1. Begin the chain interactions counter $j = 1$ and give initial values $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$ to $(\boldsymbol{\beta}, \boldsymbol{\gamma})'$.
2. Move the vector $\boldsymbol{\beta}$ to a new value $\boldsymbol{\psi}$ generated from the proposed density $q_1(\boldsymbol{\beta}^{(j-1)}, .)$.
3. Calculate the acceptance probability of movement, $\alpha(\boldsymbol{\beta}^{(j-1)}, \boldsymbol{\psi})$. If the movement is accepted, then $\boldsymbol{\beta}^{(j)} = \boldsymbol{\psi}$. If it is not accepted, then $\boldsymbol{\beta}^{(j)} = \boldsymbol{\beta}^{(j-1)}$.
4. Move the vector $\boldsymbol{\gamma}$ to a new value $\boldsymbol{\psi}$, generated from the proposed density $q_2(\boldsymbol{\gamma}^{j-1}, .)$.
5. Calculate the acceptance probability of movement, $\alpha(\boldsymbol{\gamma}^{(j-1)}, \boldsymbol{\psi})$. If the movement is accepted, then $\boldsymbol{\gamma}^{(j)} = \boldsymbol{\psi}$. If it is not accepted, then $\boldsymbol{\gamma}^{(j)} = \boldsymbol{\gamma}^{(j-1)}$.
6. Finally, change the counter from j to $j + 1$ and go to 2 until the convergence is reached.

4.2. Bivariate beta regression models

In the bivariate beta regression models, samples of the regression parameter models are obtained from the Bayesian algorithm proposed in section 4.1.

To get samples from the posterior parameter distribution for θ we propose the reparametrization $\alpha = \log\left(\frac{1-\theta}{1+\theta}\right)$, taking into account that $-1 < \theta < 1$. Thus, we assume a normal prior distribution for α and get samples of θ from the posterior samples of α obtained by a random walk.

5. Simulation study

In this simulation, we assume a bivariate beta regression model with mean and dispersion models given by

$$\text{logit}(\mu_{1i}) = 1.0 - 1.0x_{1i} + 0.2x_{2i}, \quad \log(\phi_{1i}) = 1 + 0.1z_{1i}, \tag{25}$$

$$\text{logit}(\mu_{2i}) = 2.0 - 1.0x_{1i} + 0.5x_{2i}, \quad \log(\phi_{2i}) = 1 - 0.1z_{1i}. \tag{26}$$

and dependence parameter $\theta = 0.5$. For each of the explanatory variables, 200 independent observations were generated from a uniform distribution $U(0, 10)$. Then, a sample from the variable of interest (Y_{1i}, Y_{2i}), was obtained from the bivariate beta distribution $B_{biv}(\mu_{1i}, \mu_{2i}, \phi_{1i}, \phi_{2i}, \theta)$ as in Trivedi and Zimmer (2007), following the next steps.

1. Independent random vectors V_1 and V_2 are generated from uniform distributions U(0,1).

2. Consider $U_1 = V_1$, $u_2 = 2v_2/\sqrt{B} - A + 1$, where $A = \theta(2u_1 - 1)$, $B = (1 - A)^2 + 4Av_2$.

3. $Y_1 = F_1(U_1; \mu_1, \phi_1)$ and $Y_2 = F_1(U_2; \mu_2, \phi_2)$.

With the structural generated data set and assuming independent normal prior distributions (with mean zero and variance 10^k, with $k = 3$) for all the parameters, a posterior sample of size 7000 was generated. For all the parameters and in all the three posterior samples generated, the chains showed a good behavior: for each of the parameters, the chains generated from different initial values showed the same shape behavior with a small transient period, a strong indication of the convergence to the posterior marginal distribution. The algorithm also showed not to be very sensible to initial values.

The posterior inferences were developed from the posterior sample obtained after a bourn-in period of 1000 initial values of the posterior chains, choosing a value every ten to have an approximately uncorrelated sample. The posterior chain samples and their respective histograms are shown in Figure 1 for mean parameters, in Figure 2 for the dispersion parameters and in Figure 3 for the dependence parameter.

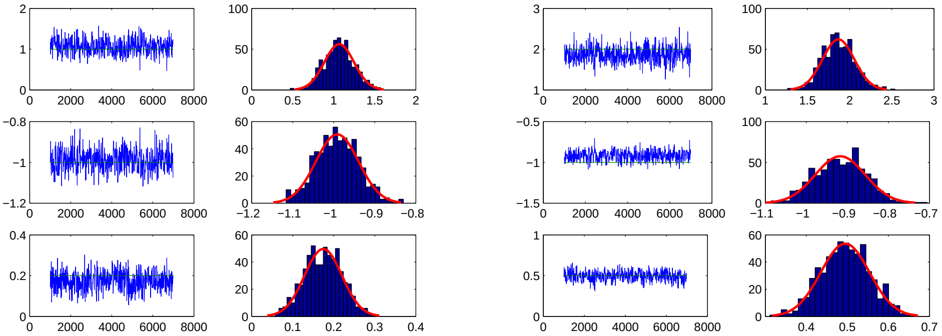

Figure 1: Posterior chain samples for the mean regression parameters of model (25) on the left, and model (26) on the right.

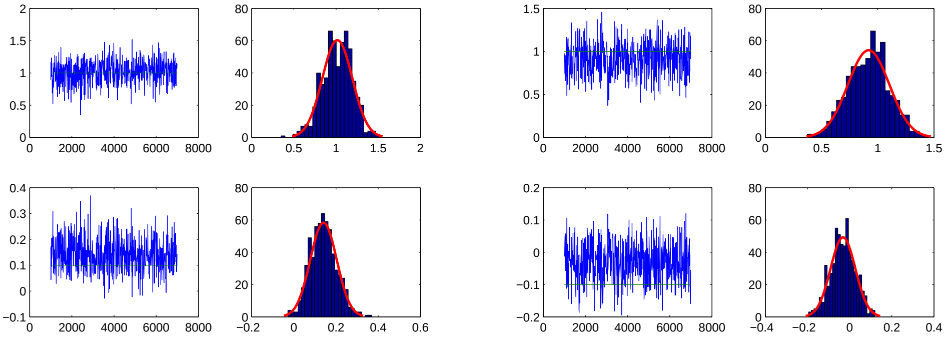

Figure 2: Posterior chain samples for the precision regression parameters of model (25) on the left, and model (26) on the right.

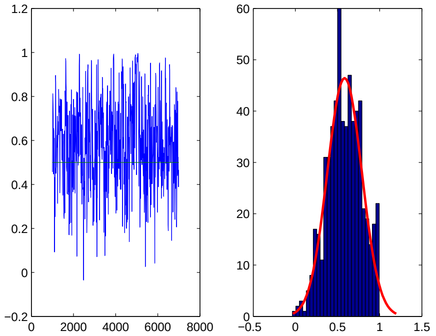

Figure 3: Posterior chain samples for the dependence parameter.

The Monte Carlo estimates of the posterior means and their respective standard deviations, obtained from the generated posterior sample, are given in Table 1, where it is possible to see that all the estimates are close to the respective true values of the parameters and all of them have small standard deviations.

		β_0	β_1	β_2	γ_0	γ_1	θ
Model 1	t.v.	1	-1	0.2	1	0.1	
	s.v.	1.0692 (0.1804)	-0.9935 (0.0513)	0.1761 (0.0461)	1.0193 (0.1836)	0.1387 (0.0648)	-
Model 2	t.v.	2	-1	0.5	1	-0.1	
	s.v.	1.8622 (0.1946)	-0.9156 (0.0605)	0.4945 (0.0585)	0.9144 (0.1858)	-0.0307 (0.0592)	-
	-	-	-	-	-	-	0.5778 (0.1968)

Table 1: Bayesian estimates of the parameters simulation study (standard deviation in parenthesis) and $\theta = 0.5$ (t.v.=true values; s.v.=simulated values).

6. An application with educational data

In this application, the random variable of interest is the average development in mathematics and language (by departments) of third year students of secondary school in Colombia. The data set was obtained from ICFES, (National Institute of Evaluation) and DANE (National Administrative Department of Statistics). The average development by department, in mathematic and language, take values in open intervals (a_m, b_m) and (a_l, b_l) and are denoted by P_m and P_l, respectively. Thus, to assume bivariate beta distribution, we define two new random variables $Y_{mi} = (P_{mi} - a_m)/(b_m - a_m)$ and $Y_{li} = (P_{li} - a_l)/(b_l - a_l)$ for the average development in mathematics and language, respectively. The explanatory variables UBN, unmet basic needs, and PORC, percentage of teachers that have postgraduate level of education, were obtained from the National Administrative Department of Statistics (DANE).

In a first approximation, we assume the model with regression models given by

$$\text{logit}(\mu_{ki}) = \beta_{k0} + \beta_{k1} UBN_i + \beta_{k2} PORC_i, \tag{27}$$

$$\log(\phi_{ki}) = \gamma_{10} + \gamma_{k1} UBN_i + \gamma_{k2} PORC_i, \;\; \text{k=1,2} \tag{28}$$

where equation (27) and (28) considering k=1,2 are related to the first and second components of the bivariate beta distribution, the mean and the variance of mathematics and languages performance, respectively. Assuming independent normal prior distribution (with mean zero and variance 10^k, with $k = 4$) for the parameters, posterior samples of size 7000 were generated in each case. The posterior inferences were developed from the posterior sample obtained after a bourn-in period of 1000 initial values of the posterior chains, choosing a value every ten to have an approximately uncorrelated sample.

	β_0	β_1	β_2	γ_0	γ_1	γ_2	θ
Math.	0.3393 (0.1229)	−0.2103 (0.1565)	0.2070 (0.2666)	4.1921 (0.0832)	1.8606 (0.1007)	1.2560 (0.1803)	0.0253 (0.3214)
Lang.	0.5374 (0.1120)	−0.2381 (0.1447)	0.2922 (0.2332)	4.4807 (0.0998)	0.7229 (0.0821)	3.0362 (0.3178)	

Table 2: Bayesian estimates of the parameters.

From Table 2 we can see that in the language case the estimate of the parameter associated with UBN is negative, an expected result given that language development is in general associated to socioeconomic factors, with unsatisfied basic needs. This is also an expected and known result for Colombia since there are several different regions, all of them having cultural and economic differences that, in general, are associated to several educational differences in the quality of the education.

The estimate of the parameter associated with the explanatory variable PERC is positive in both mathematics and language. In the educational system we expected that a better educational level of the teacher should be positively associated with better mathematical and language performance of the students.

Table 2 shows that the parameters of the precision models are different from zero as observed in the credible 95 % intervals obtained from the simulated Gibbs samples.

7. Conclusions

In this paper a bivariate beta regression model was proposed, assuming a weak dependence between the variables of interest where this dependence is modeled by a Farlie-Gumbel-Morgentern (FGM) copula function.Two examples were introduced, including a simulated study and an application with a real data set.The Bayesian methodology used to find posterior estimates of the parameters showed good performance. These results could be of great interest in applications considering bivariate beta data in the presence of covariates.The bivariate beta distribution could be generalized using other copula functions, depending on the existing dependence between the beta variables usually verified by a preliminary data analysis, that can be fitted using the same Bayesian methodology introduced in this paper.

References

[1] Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*,**2**, 244263.

[2] Branscum A. J., Johnson W. O., Thurmond M. C. (2007). Bayesian beta regression: Applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian and New Zealand Journal of Statistics*, **49**, 287-301.

[3] Bury, K. (1999). Statistical Distributions in Engineering. New York: Cambridge University Pres

[4] Cepeda, E.C. (2001). Variability Modeling in Generalized Linear Models, *Unpublished Ph.D. Thesis. Mathematics Institute, Universidade Federal do Rio de Janeiro.*

[5] Cepeda C. E. and Gamerman D.(2001). Bayesian modeling of variance heterogeneity in normal regression models, *Brazilian Journal of Probability and Statistics* , **14**,207-221.

[6] Cepeda, C.E. (2005). Factores asociados al logro cognitivo en matematicas, *Revista de Educacin*, 336, 503-514.

[7] Cepeda C. E. and Gamerman D. (2005). Bayesian methodology for modeling parameters in the two parameter exponential family. *Estadstica*, **57**, 93-105.

[8] Cepeda-Cuervo, E. and Achcar J. A. (2010). Heteroscedastic Nonlinear Regression Models *Communications in Statistics - Simulation and Computation*, **39**(2):405-419

[9] Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, Biometrika **65**, 141-151.

[10] Cribari-Neto F (2005). Improved Maximum Likelihood Estimation in a New Class of Beta Regression Models. *Brazilian Journal of Probability and Statistics*, **19**(1), 13-31.

[11] Donoso, D.S. (2002) School efficiency and socioeconomic differences: on the results of assessment exams of the Quality of Education in Chile. *Educacao e Pesquisa*, **28**(2), 11-23.

[12] Ferrari, S., Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions, *Journal of Applied Statistics* **31**, 799-815.

[13] Gelfand A. E. and Smith A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85** (410), 398-409.

[14] Gumbel, E. J. (1960). Bivariate exponential distributions, J. Amer. Statist. Assoc., **55**, 698-707

[15] Johnson, N., Kots S., Balakrishnan N.(1994). Continuous Univariate Distributions, John Wiley & Sons, Chichester, England.

[16] Jorgensen, B. (1997). Proper dispersion models (with discussion). *Brazilian Journal of Probability and Statistics*, **11**, 89-140.

[17] Nelsen, R. B. (1999), An Introduction to Copulas, New York: Springer

[18] Olkin, I. and Liu, R. (2003). A bivariate beta distribution. *Statistics and Probability Letters*, **62**, 407-412.

[19] Paolino, P., (2001). Maximun likelihood estimation of models with beta distributed diferent variables. *Political analysis*, **9**, 4, 325-346.

[20] Sancho G.J. (2006) Aprender a los 15 aos: factores que influyen en este proceso, *Revista de Educacin*, N extraordinario, 171-193.

[21] Simas A. B., Barreto-Souza W., Rocha A. V. (2010). Improved Estimators for a General Class of Beta Regression Models, *Computational Statistics & Data Analysis*, **54**(2), 348-366.

[22] Smithson M., Verkuilen J. (2006). A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables. *Psychological Methods*, **11**(1),54-71.

[23] Trivedi, P. and D. Zimmer (2007). Copula Modeling: An Introduction for Practitioners. *Foundations and Trends in Econometrics*, **1**(1), 1-110.