Revista Colombiana de Estadística
Junio 2013, volumen 36, no. 1, pp. 1 a 21

A Bayesian Approach to Parameter Estimation in Simplex Regression Model: A Comparison with Beta Regression

Un enfoque bayesiano para la estimación de los parámetros del modelo regresión Simplex: una comparación con la regresión Beta

Freddy Omar López[a]

Universidad de Valparaíso, Valparaíso, Chile

Abstract

Some variables are restricted to the open interval $(0, 1)$ and several methods have been developed to work with them under the scheme of the regression analysis. Most of research consider maximum likelihood methods and the use of Beta or Simplex distributions.

This paper presents the use of Bayesian techniques to estimate the parameters of the simplex regression supported on the implementation of some simulations and a comparison with Beta regression. We consider both models with constant variance and models with variance heterogeneity. Regressions are exemplified with heteroscedasticity.

***Key words*:** Beta distribution, Gibbs sampler, Heterogeneous, Proportions, Simplex distribution, Variances.

Resumen

Algunas variables están restringidas al intervalo abierto $(0, 1)$ y para trabajar con ellas se han desarrollado diversos métodos bajo el esquema del análisis de regresión. La mayoría de ellos han sido concebidos originalmente para ser estimados por métodos de máxima verosimilitud. Los más naturales parecen descansar especialmente sobre las distribuciones Beta o Simplex.

En este trabajo se presenta el uso de técnicas Bayesianas para la estimación de los parámetros de la regresión Simplex respaldada con la aplicación de algunas simulaciones y comparaciones con la regresión Beta. Se presentan resultados para modelos de varianza constante y de varianza heterogénea para cada individuo. Se presenta un ejemplo con datos reales.

***Palabras clave*:** distribución beta, distribución simplex, muestreador de Gibbs, proporciones, varianza heterogénea.

[a]PhD Student. E-mail: freddy.vate01@gmail.com

1. Introduction

Researchers frequently are dealing with situations where they are interested in modelling proportions, percentages or values within the open interval $(0, 1)$, according to one or several covariates, within the architecture of the regression models. This has usually been addressed with different approaches, including: linear regression, logistic regression, nonlinear regression, tobit regression, among others. However, most of them are not the natural way of working with such variables.

For this type of variable, the normal assumption, underlying in most of the mentioned techniques, it is not supported, invalidating conclusions that could be obtained from these results. Response variable's asymmetry and multicollinearity are two of the most frequent problems which the normal model cannot deal with.

In this situation, some alternatives have been developed such as Beta regression which take the general linear model advantages and the Simplex distribution, which is part of a more general class of models, the *dispersion models*.

These mentioned techniques have been developed to analyze variables that belong to the open interval $(0, 1)$ and not to $[0, 1]$. This distinction has been made by Kieschnick & McCullough (2003) in a comparative study as other authors. They recommended to use the Beta distribution or a quasi-likelihood based model when it is required to work with this type of variable.

As a comment to Paolino (2001), Buckley (2003) used the Bayesian paradigm to estimate the parameters from a Beta regression through the Metropolis-Hasting algorithm with non-informative previous distributions. This model contemplates the posibility to manage the heterogenity, besides the mean, by using two submodels corresponding to the location and dispersion submodels (Smithson & Verkuilen 2006). The research done by Paolino (2001) originally used a maximum likelihood method to estimate parameters. Ferrari & Cribari-Neto (2004) also apply this method.

Song, Qiu & Tan (2004) developed a similar model considering two submodels (one for a location parameter and another for a dispersion parameter) with a response simplex variable. The method to estimate the parameters by these authors was the generalized estimating equations (GEE).

In this work we consider a Bayesian approach for the estimation of the regression parameters and some simulations using the Gibbs sampler. Previous distributions to regression parameters have been normal with a high variance. Also, the estimation methods will be applied to a real dataset.

The main purpose of this work is to present the estimation by Bayesian methods of the simplex regression's parameters. Additionally, since Beta regression has the same objective of modelling proportions and rates, both methods will be compared some datasets generated by one or the other underlying model. We will be make emphasis on the details of the simplex distribution given the fact that the features of the beta distribution enjoy more fame in the literature than the simplex model.

This paper is structured as follows: in the Section 2 we present the simplex distribution, simplex regressions and the estimation method used in this investigation. Also, the beta regression and the comparison strategy in order to compare both models. In Section 3 we present some simulations and an application to real dataset. Finally in Section 4 some conclusions about this work.

2. Regression Models

2.1. Dispersion Models and Simplex Distribution

The simplex distribution is a distribution that belongs to the family of *dispersion models*, with location and dispersion parameters μ and σ^2, respectively (also abreviatted as $\mathsf{DM}(\mu, \sigma^2)$).

The *exponential dispersion* family density (ED) has the form

$$p(y; \theta, \phi) = \exp\left\{\frac{y\theta - \kappa(\theta)}{a(\theta)} + C(y, \phi)\right\},\ y \in \mathcal{C} \tag{1}$$

for some functions $a(\cdot), \kappa(\cdot)$ y $C(\cdot)$ with parameters $\theta \in \Theta$ and $\phi > 0$ and $\mathcal{C}$ is the support of the density. In particular, it is known that κ is the cumulant generating function. Note that ED is the classical *exponential family* of the random component in the GLM framework.

The general form of a dispersion model is

$$p(y; \mu, \sigma^2) = a(y; \sigma^2) \exp\left\{-\frac{1}{2\sigma^2} d(y; \mu)\right\},\ y \in \mathcal{C} \tag{2}$$

where $\mu \in \Omega, \sigma > 0$ and $a \geq 0$ is a normalizer term, independent of μ. Function d is known as the *unit deviance* and is defined in $(y, \mu) \in (\mathcal{C}, \Omega)$ and it must satisfy some additional properties (Song 2007).

A simple advantage over the classical exponential family parametrization in (1) is that both, mean and dispersion parameters, μ and σ^2, are explicitly in the density expression (2) whereas in (1), $\mu = E(Y) = \kappa'(\theta)$.

More precisely the parameter $\mu = E(Y)$ and $\mathrm{Var}(Y) = \frac{\sigma^2}{V(\mu)}$, where $V(\mu)$ is directly related with $d(\cdot;\cdot)$, i.e.

$$V(\mu) = \frac{2}{\left.\frac{\partial^2 d(y;\mu)}{\partial \mu^2}\right|_{y=\mu}},\ \mu \in \Omega$$

This function is known as the "unit variance function".

Specifically, if y follows a simplex distribution, that is $y \sim S^-(\mu; \sigma^2)$, then (2) takes the form

$$p(y; \mu, \sigma^2) = [2\pi\sigma^2\{y(1-y)\}^3]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2} d(y; \mu)\right\},\ y \in (0,1),\ \mu \in (0,1) \tag{3}$$

In particular, where

$$a(y;\sigma^2) = [2\pi\sigma^2\{y(1-y)\}^3]^{-\frac{1}{2}}$$

and

$$d(y;\mu) = \frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2},\ y \in (0,1),\ \mu \in (0,1)$$

It follows that $\mathsf{E}\{d(Y;\mu)\} = \sigma^2$, $\mathsf{E}\{d'(Y;\mu)\} = 0$, $\mathsf{Var}\{d(Y;\mu)\} = 2(\sigma^2)^2$. These and others features can be studied in detail at Song (2007). Other inferential properties can be studied in the seminal paper by Barndorff-Nielsen & Jørgensen (1991).

The distribution can have one or two modes and can take the approximate shape of a bell, U, J, or L (also known as reverse-J) for different combinations of its parameters. It is important to note that the simplex distribution cannot emulates a flat distribution as the uniform distribution on the interval $(0,1)$.

Figure 1 presents several examples: simplex distributions with mean values: $0.1, 0.25, 0.50, 0.75$ and 0.90 with different dispersion parameters. Note that when the second parameter is increased, the curves are becoming flatter.

2.2. Simplex Regression Model

2.2.1. Introduction

Let be $Y_1,\dots,Y_n$ independent random variables following the distribution in equation (3) with mean μ_i and dispersion parameter σ_i^2, and let be $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $\mathbf{w}_i = (w_{i1}, w_{i2}, \dots, w_{iq})$, $i = 1,\dots,n$, vectors of covariate information. It is important to note that covariables $\mathbf{x}$ and $\mathbf{w}$ can be identical or they could be subsets of each other. We want to model the mean value μ_i and the dispersion parameter σ_i^2.

Similar to Cepeda & Gamerman (2001), Smithson & Verkuilen (2006) and Song et al. (2004), two link functions, g and h will be considered one for each parameter in the simplex distribution.

A convenient function g for the mean is the logit function, which ensures the parameter μ is in the open interval $(0,1)$. More specifically

$$g(\mu_i) = \log\frac{\mu_i}{1-\mu_i} = \mathbf{x}_i^\top\boldsymbol{\beta} \tag{4}$$

where $\boldsymbol{\beta} = (\beta_0,\dots,\beta_p)$ is a vector of unknown parameters. Equation (4) is also known as the *location submodel.*

The logit function has an extensive application in the statistic field. This transformation helps to give answers in terms of the *odds ratio*. This is because the odd ratio between the predictive variable and its response variable can be found by using the relation $\mathsf{OR} = \exp(\beta_k)$, $k = 1,\dots,p$.

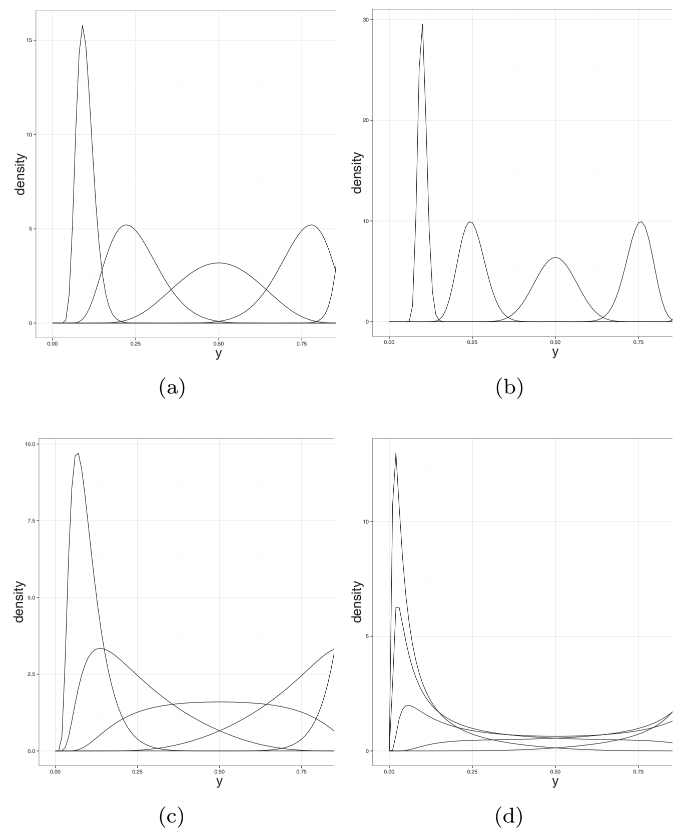


FIGURE 1: Different shapes for the simplex distribution. The distributions have as the mean value parameter $\mu = 0.1, 0.25, 0.5, 0.75, 0.9$ and different values for dispersion. (a) $\sigma = 1$; (b) $\sigma = 0.5$; (c) $\sigma = 2$ and (d) $\sigma = 5$.

On the other hand, the dispersion parameter σ_i^2 must be positive and a function h that enjoys this property is the logarithm function. So

$$h(\sigma_i^2) = \log(\sigma_i^2) = \mathbf{w}_i^\top \boldsymbol{\delta} \tag{5}$$

where $\boldsymbol{\delta} = (\delta_0, \dots, \delta_q)$ is a vector of unknown parameters that must be estimated. The equation (5) is known as the *dispersion submodel*.

2.2.2. Parameter Estimation

Maximum Likelihood

The classical theory of maximum likelihood estimation for the exponential family models (McCullagh & Nelder 1989) is very related with the maximum likelihood estimation for dispersion models as a special case. In the specific case of simplex distribution and the general linear model the score equation (derivative of the likelihood with respect to parameters) is a given by

$$\sum_{i=1}^{n} \mathbf{x}_i \{\mu_i(1-\mu_i)\} \delta(y_i;\mu_i) = 0 \tag{6}$$

where

$$\delta(y;\mu) = \frac{y-\mu}{\mu(1-\mu)} \left\{ d(y;\mu) + \frac{1}{\mu^2(1-\mu)^2} \right\}$$

Equation (6) is solved using Newton-Raphson or quasi-Newton algorithm.

In particular, it is necessary to introduce an estimation of the dispersion parameter σ^2. In this situation it is common to replace σ^2 with

$$\widehat{\sigma}^2 = \frac{1}{(n-p+1)} \sum_{i=1}^{n} d(y_i;\widehat{\mu}_i)$$

Interested readers are referred to Jørgensen (1997) and Song (2007) for more details. In this paper the maximum likelihood method is not considered.

Markov Chain Monte Carlo Sampling

With the aim of estimating the parameters of equations (4) and (5), we specify the likelihood function

$$L(\boldsymbol{\beta},\boldsymbol{\delta}) = \prod_{i=1}^{n} a(y_i; h^{-1}(\mathbf{w}_i^\top \boldsymbol{\delta})) \exp\left\{ -\frac{1}{2h^{-1}(\mathbf{w}_i^\top \boldsymbol{\delta})} d(y_i; g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})) \right\} \tag{7}$$

which posterior distribution is expressed as

$$p((\boldsymbol{\beta},\boldsymbol{\delta}) \mid \mathbf{y}) \propto L(\boldsymbol{\beta},\boldsymbol{\delta}) p(\boldsymbol{\beta},\boldsymbol{\delta}) \tag{8}$$

where $p(\boldsymbol{\beta},\boldsymbol{\delta}) = p(\boldsymbol{\beta})p(\boldsymbol{\delta})$ are the previous distribution of parameters under the assumption that they are independent to each other. In this work it is assumed that each parameter $\beta_i, i = 1, \ldots, p$ and $\delta_j, j = 1, \ldots, q$ follow a non informative distribution centered at 0 and a large variance (about $1,000$). With this information, it is possible to use several Bayesian mechanisms in order to estimate the parameters. We have chosen a Gibbs sampling approach due to because the relative ease to be implemented.

In order to define the Bayesian regression modelling framework, we specify

$$\begin{aligned} &y_i \mid \mu_i, \sigma_i^2 \sim S^-(\mu_i, \sigma_i^2) \\ &g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} \\ &h(\sigma_i^2) = \mathbf{w}_i^\top \boldsymbol{\delta} \end{aligned} \tag{9}$$

It is important to note that the models in this section are applicable to response variables y which range strictly in the open interval $(0, 1)$. However, in some situations, it is possible to have data where $y = 0$ or $y = 1$ (for instance, it can be the case where none person support the candidate's management; or that 100% of individuals under observation in a clinical trial have had reacted positively to certain stimuli). This situation can be addressed with different strategies. One of them is to replace all values 0 by a very small quantity $\epsilon > 0$ and all 1 values by $1 - \epsilon$ respectively. In other situations, when the theorical maximum and minimum values, β and α, are known the followings can be used

$$y^{\mathsf{new}} = \frac{(n-1)(y-\alpha)}{(\beta-\alpha)n} + \frac{1}{2n} \tag{10}$$

where n is the length of y. These approximations have been considered in the context of Beta regression by Smithson & Verkuilen (2006), Zimprich (2010), Verkuilen & Smithson (2011) and Eskelson, Madsen, Hagar & Temesgen (2011). This approach is not considered in this work.

2.3. Comparison to the Beta Regression Model

Beta regression has been studied with much interest on the last years (Ferrari & Cribari-Neto 2004, Ospina & Ferrari 2010, Cribari-Neto & Zeileis 2010, Cepeda & Garrido 2011, Cepeda 2012). In order to model proportions and rates.

The probability density function of a Beta distribution is given by

$$p(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1},\ 0 < y < 1$$

where Γ is the gamma function.

Considering $\mu = \frac{p}{p+q}$ and $\phi = p + q$ this produces $p = \mu\phi$ and $q = (1-\mu)\phi$. This will be the parametrization used in this work. A different parametrization based on mean and variance is studied by Cepeda (2012).

The shape of this distribution could have a variety of options. At most, it could have a single mode or a single antimode; it can show a bell-shaped, J and L-shaped and, among its particular cases, are the triangular distribution, uniform distribution and power function distribution (Johnson, Kotz & Balakrishnan 1994).

Beta regression is the most adequate model to be compared to the simplex regression because it is possible to model individual dispersion on the data (Cribari-Neto & Zeileis 2010).

It has been estimated traditionally using maximum likelihood methods but also Bayesian methods (Buckley 2003, Branscum, Johnson & Thurmond 2007, Cepeda & Garrido 2011, Cepeda 2012). In this work Bayesian methods will be used in order to estimate the parameters for Simplex and Beta regressions.

2.4. Model Comparison

2.4.1. Deviance Information Criterion

A way to compare models from the Bayesian perspective is through the DIC measure (Spiegelhalter, Best, Carlin & van der Linde 2002, Gelman, Carlin, Stern & Rubin 2003). This measure uses the *deviance* which is defined in its general form as

$$D(y, \theta) = -2 \log p(y|\theta)$$

where $p(y \mid \theta)$ is the likelihood of the data and θ are the parameters of the model. This measure depend both upon θ as y.

A measure which depend only of data y is $D_{\widehat{\theta}}(y) = D(y, \widehat{\theta}(y))$, which uses a point estimator of θ and is computed from simulations. The average over the posterior distribution is given by $D_{\text{avg}} = \mathrm{E}(D(y, \theta) \mid y)$, whose estimator is

$$\widehat{D}_{\text{avg}}(y) = \frac{1}{n} \sum_{i=1}^{n} D(y, \theta^i)$$

Another important measure, known as the *effective number of parameters* is defined as

$$p_D = \widehat{D}_{\text{avg}}(y) - D_{\widehat{\theta}}(y)$$

Finally, the *deviance information criterion* (DIC) is defined by

$$\text{DIC} = 2\widehat{D}_{\text{avg}}(y) - D_{\widehat{\theta}}(y)$$

with smaller values suggesting a better-fitting model.

2.4.2. Comparison of Ordered Simulated Data Against Ordered Observed Data

A strategy to compare the performance of the models is simulate replicated data y^{rep}, and compare it with the real data, y. The comparison can be done ordering the simulated values, $y_{(i)}^{\text{rep}}$, and displaying it against the real ordered data, $y_{(i)}$. If at the moment of plotting, they are close to an identity function, then we have evidences of a good model. Moreover, we can appreciate values that can be outliers.

To create simulated data, y^{rep}, samples are taken following a model with the parameters $\widehat{\theta}$, estimated using real data (in this case, it will be sampled from Simplex and Beta distribution). To gain precision, it is usual to simulate several datasets and at the moment of plotting, to display empirical confidence intervals for each point of the observed data $y_{(i)}$.

3. Data Analysis

The following sections will show the performance of the simplex and beta regression. The simulation was followed using a similar scheme like the one by Song et al. (2004).

In each Section of 3.1 two types of dataset will be simulated. One, keeping a constant dispersion and another varying the dispersion cross the individuals. In Section 3.1.1 all data follow a simplex distribution and simplex and beta models are considered. In a similar way, data in the Section 3.1.2 lie under a beta distribution and the models used to these data are beta and simplex.

All simulations and computations were done using the `R` software (R Development Core Team 2011). Bayesian estimation was done using the Gibbs sampling using the `R2OpenBUGS` and `rjags` libraries (Sturtz, Ligges & Gelman 2005, Martyn 2011). All chains have the minimum requirements to think they have converged (i.e. Geweke diagnostic, Gelman-Rubin diagnostic, autocorrelation).

3.1. Simulation Study

3.1.1. Simulating Simplex Data

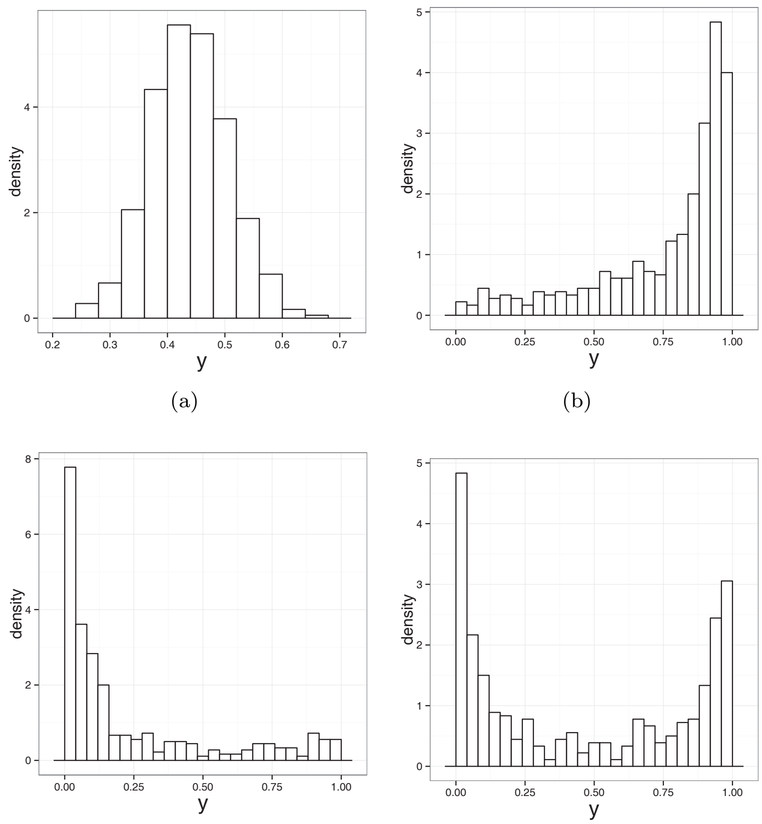
Firstly 450 independent observation $y_i, i = 1, \ldots, 450$ were obtained, belonging to a Simplex distribution with parameters (μ_i, σ^2) with the following specifications

$$\begin{cases} \operatorname{logit}(\mu_i) = \beta_0 + \beta_1 T_i + \beta_2 S_i \\ \log(\sigma^2) = \delta_0 \end{cases} \tag{11}$$

where the variable $T \in \{-1, 0, 1\}$ emulates the level of some drug and $S \in \{0, \ldots, 6\}$ suggests the illness severity. To each level of T 150 individuals were taken and from S a random sample based on a binomial distribution was taken with parameters $n = 7$ y $p = 0.5$.

Parameters of equation (11) have been fixed to emulate various shapes of y (for instance: bell-shaped, `J`, `L`, `U`). Some of these shapes are plotted on figure 2.

After applying the model strategy in (9) the results can be appreciated in Table 1 and some of its realizations can be seen in Figure 3. All parameters were estimated with a four-chain run of 30,000 iterations length. Four chains of 30,000 length each were estimated and there its first 15,000 values were discarded from each one of them. It is important to note that in general, simplex estimation of parameters is close to real values, however, it seems there is a tendency when δ_0 increases then β_j, $j = 0, 1, 2$ are distant from real values. Moreover, we note that when y variable is bell-shaped then the estimated location parameters using beta or simplex model are very similar. Coefficients marked with a † symbol means that its Bayesian confidence interval includes the 0 value.



(a) (b) (c) (d)

Figure 2: Simulations under homogeneous simplex models: (a) Bell-shaped ($\beta_0 = 0.1$, $\beta_1 = -0.1$, $\beta_2 = 0.1$, $\sigma = 0.5$); (b) J-shaped ($\beta_0 = -0.5$, $\beta_1 = 0.5$, $\beta_2 = -0.5$, $\sigma = \sqrt{15}$); (c) L-shaped ($\beta_0 = 0.1$, $\beta_1 = -0.1$, $\beta_2 = 0.1$, $\sigma = 0.5$); and, (d) U-shaped ($\beta_0 = 0.1$, $\beta_1 = -0.1$, $\beta_2 = 0.1$, $\sigma = 0.5$).

Additionally, DIC measures suggests both models are very competitive. Values estimated for the location submodels reach the greatest differences from real values when the shape of data y have form of U; in all cases the parameter of dispersion was estimated with high precision.

Second, several models were estimated varying the dispersion submodel according to the following specifications

$$\begin{cases} \operatorname{logit}(\mu_i) = \beta_0 + \beta_1 T_i + \beta_2 S_i \\ \log(\sigma_i^2) = \delta_0 + \delta_1 T_i \end{cases} \tag{12}$$

where the parameters value β_j, $j = 0, 1, 2$ have been kept as in the previous exercise and δ_j, $j = 0, 1$ have been varied as shows Table 2 to preserve shapes similar to those shown in Figure 2.

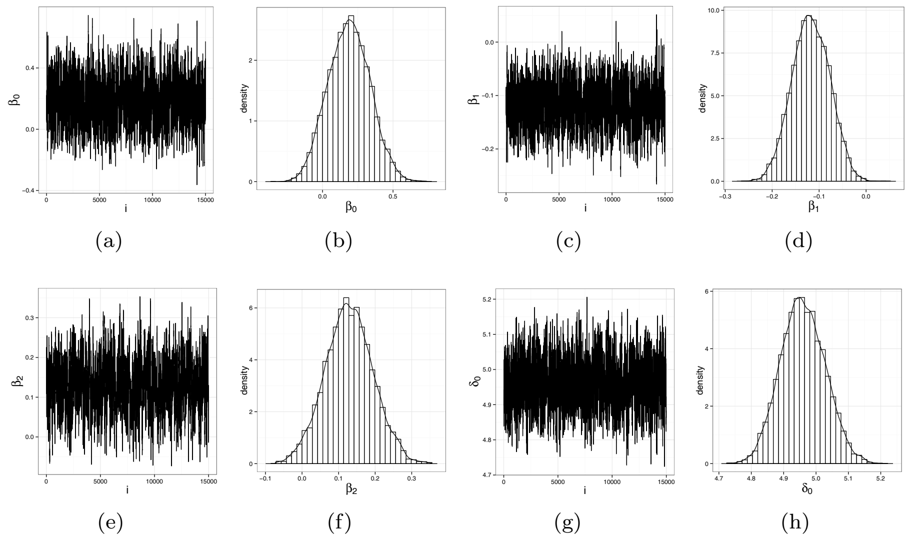


FIGURE 3: Simulation of some chains for the homogeneous simplex model with U shaped: $\beta_0 = 0.1$, $\beta_1 = -0.1$, $\beta_2 = 0.1$, $\sigma^2 = 150$: (a) and (b) are summaries for parameter β_0; (c) and (d) for β_1; (e) and (f) for β_2; (g) and (h) for δ_0.

In the same way, the results from a four-chain run of 30,000 iterations (15,000 burn-in) are presented in Table 2. Additionally, when the shape of the distribution is like a bell, estimated parameters of the location submodel in simplex and beta model are extremely similar and according to DIC, the superiority of a model over the other is not pronounced. However, these estimated values are clearly distant from its true values. When the shape of the distribution is like a J or L then the estimated location parameters are closer to true values. Estimation of dispersion parameters were also close to its true values.

3.1.2. Simulating Beta Data

Also, several models following equations (11) and (12) were considered where the support distribution is beta. The structures were estimated with beta and simplex models and results are shown in Tables 3 y 4.

It can be appreciated in Table 3 that in some cases, when beta distribution is bell-shaped, some estimations (beta and simplex) tend to be similar in its location submodel. The beta estimation seems, however, to be more distant from its true parameters values; for instance, when the distribution has shape of U given that most of its location parameters include the 0 value inside its empirical highest posterior density.

The heterogeneous case (see Table 4) was not very different. Estimated parameters are more distant from its true values in most of the cases (shapes). In several of them, the DIC measure point out that the preferred model is the simplex one.

Table 1: Homogeneous simplex models: Results after fitting Simplex and Beta regression models.

Bell − shaped								
	Simplex	Beta		Simplex	Beta		Simplex	Beta
β_0 (0.1)	0.13	0.13	β_0 (0.1)	0.10	0.10	β_0 (0.1)	0.11	0.10
β_1 (-0.1)	-0.11	-0.11	β_1 (-0.1)	-0.10	-0.10	β_1 (-0.1)	-0.10	-0.10
β_2 (0.1)	0.10	0.10	β_2 (0.1)	0.10	0.10	β_2 (0.1)	0.11	0.11
$\delta_0(\log 0.1)$	-2.32	10.30	$\delta_0(\log 0.01)$	-4.64	14.93	$\delta_0(\log 0.25)$	-1.45	8.59
DIC	-1673	-1677	DIC	-2712	-2713	DIC	-1293	-1290
J								
	Simplex	Beta		Simplex	Beta		Simplex	Beta
β_0 (-0.5)	-0.44	-0.45	β_0 (-0.5)	-0.52	-0.34	β_0 (-0.5)	-0.60	$-0.23^\dagger$
β_1 (0.5)	0.49	0.49	β_1 (0.5)	0.51	0.43	β_1 (0.5)	0.54	0.38
β_2 (-0.5)	-0.51	-0.49	β_2 (-0.5)	-0.46	-0.40	β_2 (-0.5)	-0.59	-0.44
$\delta_0(\log 1)$	$0.03^\dagger$	6.90	$\delta_0(\log 5)$	1.60	4.19	$\delta_0(\log 15)$	2.77	2.63
DIC	-1298	-1142	DIC	-748.40	-611	DIC	-685	-481
L								
	Simplex	Beta		Simplex	Beta		Simplex	Beta
β_0 (0.5)	0.60	0.56	β_0 (0.5)	0.45	0.25	β_0 (0.5)	0.37	$0.04^\dagger$
β_1 (-0.5)	-0.53	-0.51	β_1 (-0.5)	-0.49	-0.41	β_1 (-0.5)	-0.45	-0.32
β_2 (0.5)	0.47	0.44	β_2 (0.5)	0.47	0.42	β_2 (0.5)	0.48	0.37
$\delta_0(\log 1)$	$0.00^\dagger$	6.91	$\delta_0(\log 5)$	1.62	4.48	$\delta_0(\log 15)$	2.72	2.74
DIC	-1246	-1123	DIC	-815	-699	DIC	-598	-457
U								
	Simplex	Beta		Simplex	Beta		Simplex	Beta
β_0 (0.1)	0.46	0.39	β_0 (0.1)	0.19	$0.13^\dagger$	β_0 (0.1)	0.18	$0.06^\dagger$
β_1 (-0.1)	-0.22	-0.18	β_1 (-0.1)	-0.11	$-0.07^\dagger$	β_1 (-0.1)	-0.12	$-0.08^\dagger$
β_2 (0.1)	0.10	$0.12^\dagger$	β_2 (0.1)	0.13	$0.09^\dagger$	β_2 (0.1)	0.13	0.12
$\delta_0(\log 50)$	3.89	0.58	$\delta_0(\log 100)$	4.51	0.08	$\delta_0(\log 150)$	4.96	$-0.31^\dagger$
DIC	-300	-125	DIC	-386	-201	DIC	-688	-385

3.2. Example with Real Data

In this section we study the relationship between the amount people of in poverty and the government form they have elected in some geographical region. We want to determine if some variables, traditionally indicators of poverty (number of people indeed poverty, suicide rate, Human Development Index) are associated with a political option in electoral preferences terms.

The relationship between these variables has been studied previously. For instance, it is documented that for some countries, suicide rates increases when a specific political party is in the government. Blakely & Collings (2002) commented that “suicide rates were indeed higher during periods of conservative government” for the investigation done with Australian data carried out by Page, Morrell & Taylor (2002). Shaw, Dorling & Smith (2002) analyzed data from England and Wales and reached similar conclusions to the point to add the subtitle to their investigation: *Do conservative governments make people want to die?*

Also there have been findings there exists out a significant association between general mortality and political preferences (Smith & Dorling 1996).

Data analyzed in this paper correspond to 322 of 335 municipalities in Venezuela (the position of Amazonas’ Governor and others municipalities were not available for that election date). These data were taken from the website of the National Electoral Council, (CNE 2008) and the National Statistical Office, (INE 2008).

TABLE 2: Heterogeneous simplex models: Results after fitting Simplex and Beta regression models.

Bell − shaped								
	Simplex	Beta		Simplex	Beta		Simplex	Beta
β_0 (0.1)	-0.05†	-0.06†	β_0 (0.1)	0.12†	0.12†	β_0 (0.1)	-0.03†	-0.03†
β_1 (-0.1)	-0.06	-0.06	β_1 (-0.1)	-0.10	-0.10	β_1 (-0.1)	-0.07	-0.07
β_2 (0.1)	0.07†	0.07†	β_2 (0.1)	0.12	0.12	β_2 (0.1)	0.06	0.06
$\delta_0(1)$	1.06	4.13	$\delta_0(0.1)$	0.15	5.57	$\delta_0(0.3)$	0.29	5.33
$\delta_1(1)$	1.19	-1.90	$\delta_1(0.1)$	0.03†	-0.13†	$\delta_0(0.2)$	0.20	-0.36
DIC	-394	-374	DIC	-639	-634	DIC	-588	-584
J								
	Simplex	Beta		Simplex	Beta		Simplex	Beta
β_0 (-0.5)	-0.42	-0.41	β_0 (-0.5)	-0.71	-0.48	β_0 (-0.5)	-0.47	-0.05†
β_1 (0.5)	0.49	0.48	β_1 (0.5)	0.54	0.45	β_1 (0.5)	0.49	0.34
β_2 (-0.5)	-0.45	-0.45	β_2 (-0.5)	-0.49	-0.52	β_2 (-0.5)	-0.57	-0.59
$\delta_0(1)$	0.99	5.58	$\delta_0(2)$	2.01	3.95	$\delta_0(3)$	3.01	2.65
$\delta_1(1)$	0.99	-2.29	$\delta_1(1)$	1.01	-2.15	$\delta_0(1)$	1.07	-2.00
DIC	-994	-896	DIC	-783	-647	DIC	-808	-601
L								
	Simplex	Beta		Simplex	Beta		Simplex	Beta
β_0 (0.5)	0.60	0.50	β_0 (0.5)	0.33	0.23	β_0 (0.5)	0.30	-0.03†
β_1 (-0.5)	-0.52	-0.47	β_1 (-0.5)	-0.46	-0.41	β_1 (-0.5)	-0.43	-0.30
β_2 (0.5)	0.49	0.51	β_2 (0.5)	0.52	0.55	β_2 (0.5)	0.51	0.50
$\delta_0(1)$	1.02	5.28	$\delta_0(2)$	1.99	4.08	$\delta_0(3)$	3.04	2.49
$\delta_1(1)$	1.03	-2.34	$\delta_1(1)$	1.12	-2.39	$\delta_0(1)$	1.01	-1.67
DIC	-946	-818	DIC	-782	-668	DIC	-623	-483
U								
	Simplex	Beta		Simplex	Beta		Simplex	Beta
β_0 (0.1)	0.25	0.32	β_0 (0.1)	0.13†	0.18†	β_0 (0.1)	0.26†	0.19†
β_1 (-0.1)	-0.13	-0.13	β_1 (-0.1)	-0.10	-0.09	β_1 (-0.1)	-0.13	-0.12
β_2 (0.1)	0.14	0.18	β_2 (0.1)	0.12	0.12†	β_2 (0.1)	0.07†	0.03†
$\delta_0(3)$	2.98	1.59	$\delta_0(4)$	4.04	0.58	$\delta_0(5)$	4.98	-0.22
$\delta_1(1)$	1.20	-1.36	$\delta_1(1)$	1.00	-0.89	$\delta_0(1)$	1.06	-0.75
DIC	-188	-98	DIC	-252	-142	DIC	-728	-446

The response variable is the proportion of people who support with their votes the political proposal lead by Hugo Chávez.

Several models were adjusted to these data and the results can be seen in Table 5. In this Table, three models for the two underlying distributions were considered. The first of them (m_{s_0} and m_{b_0}) are the saturated models and m_{s_1} and m_{b_1} are the null models. Searching over additive structures in function of DIC give us as best models those labeled as m_{s_2} y m_{b_2}. For both, the same variables are significant for location and dispersion submodels. Note that, in general terms, coefficients for location submodels are very similar. This can be expected because the shape of the variable `% Chávez` is symmetric (see Figure 5 (b)).

Table 3: Homogeneous beta models: Results after fitting Beta and Simplex regression models.

	Beta	Simplex		Beta	Simplex		Beta	Simplex
Bell − shaped								
β_0 (0.1)	-0.07†	-0.15†	β_0 (0.1)	0.05†	0.07†	β_0 (0.1)	0.17	0.17
β_1 (-0.1)	-0.06	-0.05†	β_1 (-0.1)	-0.09	-0.09†	β_1 (-0.1)	-0.13	-0.13
β_2 (0.1)	0.13	0.15	β_2 (0.1)	0.12	0.12	β_2 (0.1)	0.09	0.09
$\delta_0(\log 30)$	3.29	1.65	$\delta_0(\log 50)$	3.80	1.25	$\delta_0(\log 200)$	5.36	0.30
DIC	-205	-176	DIC	-286	-285	DIC	-597	-593
J								
	Beta	Simplex		Beta	Simplex		Beta	Simplex
β_0 (-0.5)	-0.09	-0.30†	β_0 (-0.5)	-0.25†	-0.99	β_0 (-0.5)	-0.20†	-0.16†
β_1 (0.5)	0.28	0.35	β_1 (0.5)	0.35	0.61	β_1 (0.5)	0.38	0.44
β_2 (-0.5)	-0.24	-0.07†	β_2 (-0.5)	-0.35	-0.27	β_2 (-0.5)	-0.42	-0.55
$\delta_0(\log 1)$	0.52	5.12	$\delta_0(\log 5)$	1.51	4.46	$\delta_0(\log 15)$	2.81	3.60
DIC	-690	-861	DIC	-533	-467	DIC	-538	-347
L								
	Beta	Simplex		Beta	Simplex		Beta	Simplex
β_0 (0.5)	0.11†	0.79	β_0 (0.5)	0.24†	0.12†	β_0 (0.5)	0.25†	0.36†
β_1 (-0.5)	-0.31	-0.59	β_1 (-0.5)	-0.40	-0.45	β_1 (-0.5)	-0.42	-0.51
β_2 (0.5)	0.24	0.34	β_2 (0.5)	0.44	0.60	β_2 (0.5)	0.40	0.46
$\delta_0(\log 1)$	0.71	5.02	$\delta_0(\log 5)$	1.85	4.59	$\delta_0(\log 15)$	2.74	3.78
DIC	-752	-949	DIC	-625	-472	DIC	-587	-399
U								
	Beta	Simplex		Beta	Simplex		Beta	Simplex
β_0 (0.1)	0.03†	-0.02†	β_0 (0.1)	0.08†	-0.01†	β_0 (0.1)	-0.13†	-0.20†
β_1 (-0.1)	-0.09†	-0.08	β_1 (-0.1)	-0.07†	-0.04†	β_1 (-0.1)	-0.03†	-0.01†
β_2 (0.1)	0.06	0.14†	β_2 (0.1)	0.12†	0.08†	β_2 (0.1)	-0.02†	-0.02†
$\delta_0(\log 1)$	0.29	5.21	$\delta_0(\log 0.5)$	-0.25	5.62	$\delta_0(\log 0.25)$	-0.60	5.81
DIC	-177	67	DIC	-352	-289	DIC	-571	-756

A sample of predicted values for all models can be appreciated in Figure 5 (a). Note that the models give a *linear* prediction, that is, crossing the approximate mean of data for each value of variable `Mortality` according to its linear nature. Both models are quite similar and its fitting is displayed in Figure 5 (a). Figures 5 (c) and (d) show the average predicted values (and its empirical error bar) for each y_i point. There were simulated 100 datasets.

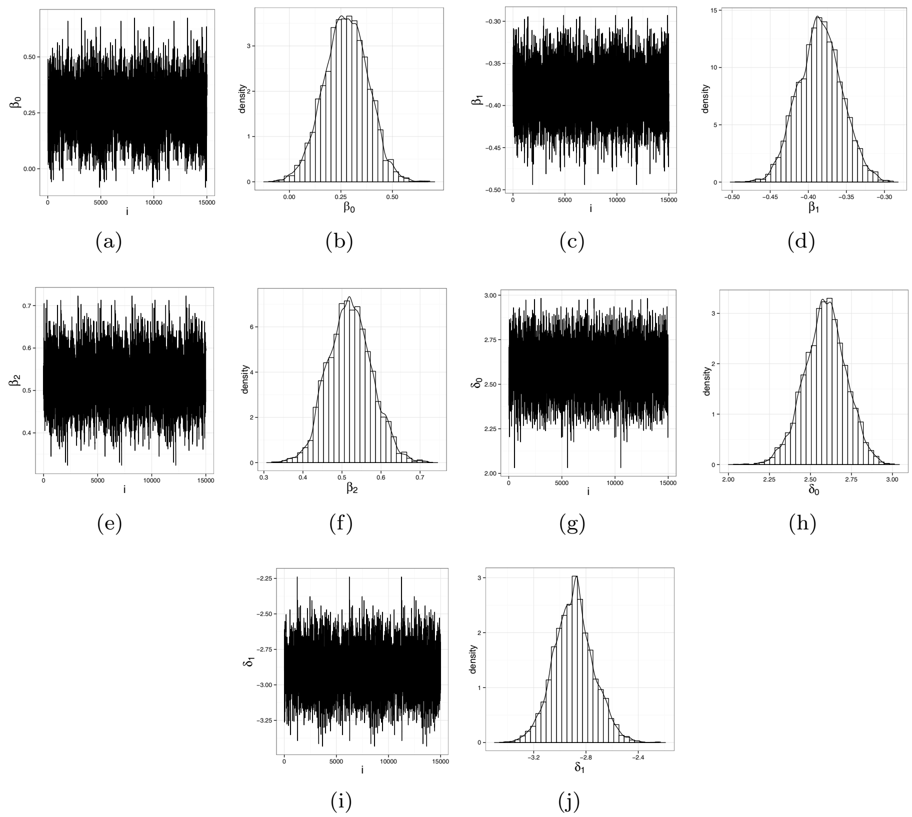


FIGURE 4: Simulation of some chains for the heterogeneous beta model with L shaped: $\beta_0 = 0.5$, $\beta_1 = 0.5$, $\beta_2 = 0.5$, $\delta_0 = 3$, $\delta_1 = 2$: (a) and (b) describes results for parameter β_0; (c) and (d) for β_1; (e) and (f) for β_2; (g) and (h) for δ_0; (i) and (j) for δ_1.

Table 4: Heterogeneous beta models: Results after fitting Beta and Simplex regression models.

	Beta	Simplex		Beta	Simplex		Beta	Simplex
Bell-shaped								
β_0 (0.1)	0.13†	0.34	β_0 (0.1)	0.08†	0.08†	β_0 (0.1)	0.11	0.11
β_1 (-0.1)	-0.11	-0.19	β_1 (-0.1)	-0.10	-0.10	β_1 (-0.1)	-0.10	-0.10
β_2 (0.1)	0.01†	0.03†	β_2 (0.1)	0.15	0.16	β_2 (0.1)	0.09	0.09
$\delta_0(3)$	2.84	2.23	$\delta_0(5)$	5.09	0.48	$\delta_0(10)$	9.97	-2.12
$\delta_1(1)$	0.97	-0.93	$\delta_1(1)$	1.08	-0.69	$\delta_1(5)$	5.12	-2.66
DIC	-152	-34	DIC	-544	-542	DIC	-1605	-1608
J								
	Beta	Simplex		Beta	Simplex		Beta	Simplex
β_0 (-0.5)	-0.17†	-0.35	β_0 (-0.5)	-0.55	-0.46	β_0 (-0.5)	-0.34	-0.41
β_1 (0.5)	0.37	0.51	β_1 (0.5)	0.44	0.41	β_1 (0.5)	0.46	0.49
β_2 (-0.5)	-0.50	-0.45	β_2 (-0.5)	-0.23	-0.09†	β_2 (-0.5)	-0.49	-0.44
$\delta_0(1)$	1.56	4.55	$\delta_0(1)$	2.29	3.98	$\delta_0(3)$	3.45	3.32
$\delta_1(1)$	0.18	-0.55	$\delta_1(5)$	2.73	-2.76	$\delta_1(2)$	1.50	-1.80
DIC	-703	-757	DIC	-884	-980	DIC	-783	-678
L								
	Beta	Simplex		Beta	Simplex		Beta	Simplex
β_0 (0.5)	-0.17†	-5.72	β_0 (0.5)	0.55	0.87	β_0 (0.5)	0.58	-3.16
β_1 (-0.5)	0.37	-1.06	β_1 (-0.5)	-0.43	-0.55	β_1 (-0.5)	-0.55	-0.58
β_2 (0.5)	-0.50	7.68	β_2 (0.5)	0.15	0.18	β_2 (0.5)	0.62	4.49
$\delta_0(1)$	1.56	36.77	$\delta_0(1)$	2.18	3.94	$\delta_0(3)$	3.27	16.11
$\delta_1(1)$	0.18	-27.85	$\delta_1(5)$	2.87	-2.64	$\delta_1(2)$	1.80	-12.65
DIC	-4121	11948	DIC	-824	-924	DIC	-1408	3936
U								
	Beta	Simplex		Beta	Simplex		Beta	Simplex
β_0 (0.1)	0.20†	-0.62	β_0 (0.1)	0.20†	0.51	β_0 (0.1)	0.01†	1.37
β_1 (-0.1)	-0.13	-0.21	β_1 (-0.1)	-0.14	-0.22	β_1 (-0.1)	-0.04†	-0.45
β_2 (0.1)	0.14	0.61	β_2 (0.1)	0.24	0.19	β_2 (0.1)	0.13	0.43†
$\delta_0(0.1)$	0.36	7.98	$\delta_0(0.1)$	0.40	4.92	$\delta_0(0.01)$	-0.11†	9.06
$\delta_1(0.1)$	0.13†	-1.97	$\delta_1(0.5)$	0.21†	-0.44	$\delta_1(0.05)$	0.02†	0.08†
DIC	-169	1358	DIC	-201	-63	DIC	-274	1450

Table 5: Parameter estimates using simplex and Bbeta regression for venezuelan election data (2008).

	Simplex model			Beta model		
	m_{s_0}	m_{s_1}	m_{s_2}	m_{b_0}	m_{b_1}	m_{b_2}
Location submodel						
Intercept	0.91	0.08	0.10	0.90	0.08	0.09
Suicides	0.03			0.02		
General Mortality	-0.10		-0.08	-0.07		-0.07
Households in poverty	-0.03			0.08		
IDH	-1.00			-1.04		
Dispersion submodel						
Intercept	-9.21	0.13	-12.23	15.53	5.84	24.31
Suicides	-0.18		-0.19	0.42		0.27
General Mortality	-0.03			-0.22		
Households in poverty	-1.69			3.44		
IDH	12.01		15.11	-13.03		-22.62
DIC	-469.03	-435.31	-476.23	-505.49	-489.87	-511.87

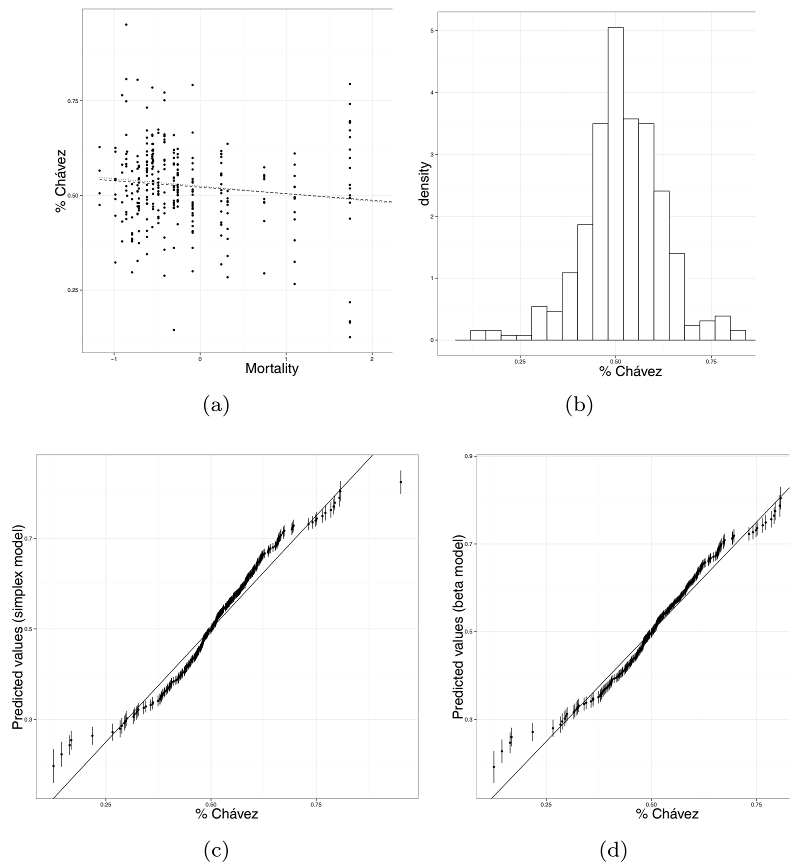


Figure 5: (a) Adjusted values for Simplex and Beta models; lines are nearly superimposed; (b) histogram of proportion of percentage of people that support Chávez; (c) ordered Chavism vs. ordered prediction based on Simplex model; (d) ordered Chavism vs. ordered prediction based on Beta model

4. Conclusions

This paper has shown how the Bayesian estimation can be applied on simplex model regression and, in addition, several simulations were performed to compare Simplex and Beta regressions. It was found that the estimation strategy produces better results when the true model is homogeneous. In particular, when the true model is homogeneous simplex, the estimates are closer to the true value parameters than the Beta model. Similar situations were found with the heterogeneous models. Most of the time, dispersion submodel parameters were estimated quite well even in the case where none parameter for the location submodel was near to its true value. Methodology was exemplified with a real dataset. For this, point estimates were pretty similar for both models: Simplex and Beta.

Further research could consider the natural extension to the (longitudinal) mixed models similar to those presented by Verkuilen & Smithson (2011) and Zimprich (2010) from the Bayesian perspective and supported by underlying simplex distribution assumption. Song et al. (2004) propose a simplex longitudinal data analysis in its marginal version.

Although, in the applications considered here, all data were inside the open interval $(0, 1)$; it is possible to model variables inside the closed interval $[0, 1]$ and there exist more adequate models such as those proposed by Cook, Kieschnick & McCullough (2008) and Ospina & Ferrari (2010).

Furthermore, it is important to investigate another alternatives for the link functions. As pointed out by Eskelson et al. (2011), the logit transformation is used because it offers an easy interpretation in terms of odds ratio but it is also possible to use the non-transformed variable. In relation with the beta regression, Giovanetti (2007) explores another alternatives to link functions and studies the empirical consequences having an incorrect specification.

In relation with Simplex regression residuals, Santos (2011) considers the situation when the parameters are estimated using the maximum likelihood method. Miyashiro (2008) proposes some diagnostic measures and performs comparisons with two real datasets estimating its parameters under Beta and Simplex assumptions. Results for those particular cases are very similar for location submodels. In that investigation, Miyashiro only studied homogeneous models using maximum likelihood.

Acknowledgments

I am in debt to Lisbeth Mora and Professor Daniel Paredes for reading the earlier stages of this research and for suggesting invaluable improvements. I also thank two anonymous referees for their comments that helped to increase the quality of this paper.

[Recibido: marzo de 2012 — Aceptado: enero de 2013]

References

Barndorff-Nielsen, O. E. & Jørgensen, B. (1991), 'Some parametric models on the simplex', *Journal of Multivariate Analysis* **39**, 106–116.

Blakely, T. & Collings, S. (2002), 'Is there a causal association between suicide rates and the political leanings of government?', *Journal of Epidemiology and Community Health* **56**(10), 722.

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 desease viruses', *Australian & New Zealand Journal of Statistics* **49**(3), 287–301.

Buckley, J. (2003), 'Estimation of models with Beta-distributed dependent variables: A replication and extension of Paolino's study', *Political Analysis* **11**, 204–205.

Cepeda, E. (2012), Beta regression models: Joint mean and variance modeling, Technical report, Universidad Nacional de Colombia.

Cepeda, E. & Gamerman, D. (2001), 'Bayesian modeling of variance heterogeneity in normal regression models', *Brazilian Journal of Probability and Statistics* **14**, 207–221.

Cepeda, E. & Garrido, L. (2011), Bayesian Beta regression models: Joint mean and precision modeling, Technical report, Universidad Nacional de Colombia.

CNE (2008), 'Consejo Nacional Electoral', `http://www.cne.gob.ve`.

Cook, D. O., Kieschnick, R. & McCullough, B. D. (2008), 'Regression analysis of proportions in finance with self selection', *Journal of Empirical Finance* **15**, 860–867.

Cribari-Neto, F. & Zeileis, A. (2010), 'Beta regression in R', *Journal of Empirical Finance* **34**(2), 1–24.

Eskelson, N. I., Madsen, L., Hagar, J. C. & Temesgen, H. (2011), 'Estimating Riparian understory vegetation cover with Beta regression and copula models', *Forest Science* **57**(3), 212–221.

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

Gelman, A., Carlin, B. P., Stern, H. S. & Rubin, D. B. (2003), *Bayes and Empirical Bayes Methods for Analysis*, 2 edn, Chapman & Hall/CRC.

Giovanetti, A. C. (2007), Efeitos da especificaçao incorreta da funçao de ligaçao no modelo de regressão beta, Master's thesis, USP, Sao Paulo.

INE (2008), 'Instituto Nacional de Estadística', `http://www.ine.gob.ve`.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1994), *Continuous Univariate Distributions*, Vol. 2, 2 edn, John Wiley & Sons.

Jørgensen, B. (1997), *The Theory of Dispersion Models*, Monographs on Statistics and Applied Probability, Taylor & Francis.

Kieschnick, R. & McCullough, B. D. (2003), 'Regression analysis of variates observed on (0,1): Percentages, proportions and fractions', *Statistical Modelling* **3**, 193–213.

Martyn, P. (2011), *rjags: Bayesian graphical models using MCMC.* R package version 3-5.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models, Second Edition*, number 37 *in* 'Monographs on Statistics and Applied Probability', London: Chapman & Hall.

Miyashiro, E. S. (2008), Modelos de regressão Beta e simplex para análise de proporçoes no modelo de regressão Beta, Master's thesis, USP, Sao Paulo.

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

Page, A., Morrell, S. & Taylor, R. (2002), 'Suicide and political regime in New South Wales and Australia during the 20th century', *Journal of Epidemiology and Community Health* **56**(10), 766–772.

Paolino, P. (2001), 'Maximum likelihood estimation of models with Beta-distributed dependent variables', *Political Analysis* **9**, 325–346.

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Santos, L. A. (2011), Modelos de Regressão Simplex: Resíduos de Pearson Corrigidos e Aplicações, PhD thesis, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.

Shaw, M., Dorling, D. & Smith, G. D. (2002), 'Mortality and political climate: How suicide rates have risen during periods of Conservative government, 1901–2000', *Journal of Epidemiology and Community Health* **56**(10), 723–725.

Smith, G. D. & Dorling, D. (1996), "'I'm all right, John": Voting patterns and mortality in England and Wales', *British Medical Journal* **313**(21), 1573–1577.

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

Song, X. K. (2007), *Correlated Data Analysis: Modeling, Analytics, and Applications*, Springer, New York.

Song, X., Qiu, Z. & Tan, M. (2004), 'Modelling heterogeneous dispersion in marginal models for longitudinal proportional data', *Biometrical Journal* **5**, 540–553.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002), 'Bayesian measures of model complexity and fit (with discussion)', *Statistical Methodology Series B* **64**(4), 583–639.

Sturtz, S., Ligges, U. & Gelman, A. (2005), 'R2WinBUGS: A package for running WinBUGS from R', *Journal of Statistical Software* **12**(3), 1–16.

Verkuilen, J. & Smithson, M. (2011), 'Mixed and mixture regression models for continuous bounded responses using the Beta distribution', *Journal of Educational and Behavioral Statistics* **000**, 1–32.

Zimprich, D. (2010), 'Modeling change in skewed variables using mixed Beta regression models', *Research in Human Development* **7**(1), 9–26.