ISSN 0104-0499

PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA
UFSCar-USP

DEs-UFSCar e SME-ICMC-USP

MODELLING MHDI WITH A MIXTURE OF SIMPLEX REGRESSION MODEL

Rosineide F. da Paz
Jorge Luis Bazán

RELATÓRIO TÉCNICO

TEORIA E MÉTODO – SÉRIE A

Abril/2017
nº 265

Modelling MHDI with a mixture of simplex regression model

Rosineide F. da Paz* & Jorge Luis Bazán[1]

Abstract

This manuscript deals with an analysis of the municipal human development index as a function of the Municipal Human Poverty Index. We propose a regression model where the response follow a mixture of simplex distribution. Estimation is performed by a Bayesian approach making use of *Gibbs sampling* algorithm.

Key words: Simplex distribution, Bayesian Analysis, Gibbs sampling, human development index,

1 Introduction

The Human Development Index (HDI) is a summary measure of long-term progress in three basic dimensions of human development that takes into account education, income and longevity indexes. The HDI is the geometric mean of normalized indexes for each of the three dimensions of human development.

The analysis of the MHDI data set as a function of the Municipal Human Poverty Index (MHPI) is presented here where the data is modeled by a mixture of two simplex distribution. The MHPI is a proportion of individuals in each city with household income equal or less than half minimum wage (R$ 255,00), August 2010 (Fundação Instituto Brasileiro de Geografia e Estatística, 2014). We consider MIDH and MHPI of the cities of Northeast region and São Paulo state in Brazil. In this region are the states of Alagoas, Bahia, Ceará, Maranhão, Paraíba, Pernambuco, Piauí, Rio Grande do Norte and Sergipe. There are 1794 cities in the Northeastern region and 645 in São Paulo state leading to a sample of size $n = 2439$.

The remainder of the manuscript is organized as follows: In Section 2 we present the Mixture of simplex regression model. The Section 3 is dedicate to Models specification and some criteria of comparison. Finally, the results are drawn in Section 4.

* *Universidade Federal & Universidade de São Paulo, São Carlos-SP, Brazil. E-mail: rfpaz@icmc.usp.br*

[1] *Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, São Carlos-SP, Brazil.*

2 Model

Let $(x_i, \boldsymbol{y}_i)$ observations where y_i represents the observed value of random variable Y_i taken value in $(0,1)$ and $\boldsymbol{x}_i = \left(\boldsymbol{x}_i^{(M)T}, \boldsymbol{x}_i^{(D)T}\right)^T$ a vector of explanatory variables with dimensions q and d, respectively, both with 1 in the first component. In addition, let's assume that Y_i is independent with density

$$f_i(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\omega}) = \sum_{j=1}^{k} \omega_j S(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_j, \boldsymbol{\delta}_j) \tag{1}$$

where

$$S(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_j, \boldsymbol{\delta}_j) = \left(2\pi\sigma_{ij}^2 \left(y_i(1-y_i)\right)^3\right)^{-1/2} \exp\left\{-\left(\frac{1}{2\sigma_{ij}^2}\right)\left(\frac{(y_i-\mu_{ij})^2}{y_i(1-y_i)\mu_{ij}^2(1-\mu_{ij})^2}\right)\right\} I_{(0,1)}(y_i)$$

is the jth component density of the mixture model given by (1), μ_{ij} and σ_{ij} is the mean and dispersion parameters, respectively, with

$$h_1(\mu_{ij}) = \boldsymbol{x}_i^{(M)T}\boldsymbol{\beta}_j \quad \text{and} \quad h_2(\sigma_{ij}) = \boldsymbol{x}_i^{(D)T}\boldsymbol{\delta}_j \tag{2}$$

$$\tag{3}$$

where h_1 and h_2 are link functions to mean and dispersion and the components of the vectors $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_k)$ and $\boldsymbol{\delta} = (\boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_k)$ are q and d-dimensional vectors of unknown regression parameters. Since the component density of the mixture model in (1) is a pdf of simplex distribution the model specified relates to a mixture of simplex regression model with k components were $\boldsymbol{\omega} = (\omega_1, .., \omega_k)$ are the weight of the mixture model.

2.1 Bayesian inference

Let's consider a unobserved random vector $Z_i = (Z_{i1}, ..., Z_{ik})$ such that $Z_{ij} = 1$ if the ith observation belongs to the jth mixture component and $Z_{ij} = 0$ otherwise, $i = 1, \ldots, n$. The augmented data likelihood to $(\mathbf{y}, \mathbf{Z})$ can be written as

$$L(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\omega}|\mathbf{y}, \mathbf{Z}) = \prod_{i=1}^{n}\prod_{j=1}^{k}\left[\omega_j S(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_j, \boldsymbol{\delta}_j)\right]^{Z_{ij}} \tag{4}$$

where $\mathbf{Z} = (Z_1, ...Z_n)$.

Assuming that y_i is assigned to component j, $Z_{ij} = 1$, the likelihood to $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$, given Z, can be write as

$$L(\boldsymbol{\beta}, \boldsymbol{\delta}|\boldsymbol{x}, \boldsymbol{y}, \mathbf{Z}, k) = \prod_{i=1}^{n}\prod_{j=1}^{k}\left[S(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}_j, \boldsymbol{\delta}_j)\right]^{Z_{ij}} \tag{5}$$

$$= \prod_{j=1}^{k} \exp\left\{-\sum_{i\in\{i:Z_{ij}=1\}}\left(\frac{(y_i-\mu_{ij})^2}{2\sigma_{ij}^2 y_i(1-y_i)\mu_{ij}^2(1-\mu_{ij})^2}\right)\right\} \prod_{i\in\{i:Z_{ij}=1\}} \left(2\pi\sigma_{ij}^2\left(y_i(1-y_i)\right)^3\right)^{-1/2}, \tag{6}$$

$$\tag{7}$$

each of these k factors can be combined with a prior distribution leading to the its full conditional posterior distribution. We specify proper prior distributions as

$$\begin{aligned}
\boldsymbol{\beta}_{lj} &\sim Normal(0, 100), \text{ for } l = 0, .., q-1 \text{ and } j = 1, ..., k, \\
\boldsymbol{\delta}_{lj} &\sim Normal(0, 100), \text{ for } l = 0, .., d-1 \text{ and } j = 1, ..., k, \\
\boldsymbol{\omega} &\sim Dirichlet(\nu_1, ..., \nu_k) \qquad (8) \\
& \qquad (9)
\end{aligned}$$

In order to simulate sample from the joint posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\omega}, Z)$ we use the Markov chain Monte Carlo approach as described in Paz *et al.* (2014).

2.2 Models specification and some criteria of comparison

We consider three mixture of simplex regression model to model the MHDI data. For the first model (M_1) we consider link function $\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$ and $log(\sigma_{ij}) = \boldsymbol{\delta}_j, j = 1, 2$ where $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j})$ and $\boldsymbol{\delta}_j = (\delta_{0j}, \delta_{1j})$. In the second model ($M_2$), link functions are considered to mean and dispersion as $\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$ and $\log(\sigma_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\delta}_j, j = 1, 2$. Finally, in the third model (M_3) we adopt link function only to the dispersion parameters as $\log(\sigma_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\delta}_j, j = 1, 2$. We assume $\boldsymbol{x}_i^T$ a vector of explanatory variables with 1 in the first component and the ith value of MHPI in the second component, $i = 1, ..., 2439$. We shall denote by M_0 the model without covariates.

The models (M_0, M_1, M_2 and M_3) was compered by estimated marginal likelihood and deviance information as expected Akaike information criteria (EAIC), expected Bayesian information criteria (EBIC) and deviance information criteria (DIC) introduced by Spiegelhalter *et al.* (2002). The MCMC output was used to approximate these criteria. The estimate of marginal likelihood was obtained based on the identity

$$m(\boldsymbol{y}) = \frac{\prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\omega}, M) p(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\omega}|M)}{p(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\omega}|\boldsymbol{y}, \boldsymbol{x}, M)} \qquad (10)$$

where $f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\omega}, M)$ is the density of ith observation to current model, M, $p(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\omega}|M)$ is the prior to the parameters and $p(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\omega}|\boldsymbol{y}, \boldsymbol{x}, M)$ is the density of posterior distribution. The approximate $p(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\omega}|\boldsymbol{y}, \boldsymbol{x}, M)$ is obtained as in Paz *et al.* (2014) where is used an approach introduced by Chib & Jeliazkov (2001) to approximate the marginal densities in the mixture models when its do not have know form. The estimate of DIC is obtained as

$$DIC = 2\bar{D} + P_D \qquad (11)$$

where $P_D = \bar{D} - \hat{D}$ with

$$\bar{D} = \sum_{g=1}^{G} \left(-2 \sum_{i=1}^{n} \log\left(f_i(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}^{(g)}, \boldsymbol{\delta}^{(g)}, \boldsymbol{\omega}^{(g)}) \right) \right) \qquad (12)$$

and

$$\hat{D} = -2\sum_{i=1}^{n} \log\left(f_i(y_i|\boldsymbol{x}_i, \bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\delta}}, \bar{\boldsymbol{\omega}})\right) \tag{13}$$

where the notation $\bar{\theta}$ main the posterior mean of the parameter θ and $\theta^{(g)}$ represent the gth estimate of the parameter θ, all the estimates are obtained from MCMC output. The EAIC and EBIC are estimated by

$$\begin{aligned} EAIC &= \bar{D} + 2 \times P \\ EBIC &= \bar{D} + 10 \times \log n \end{aligned} \tag{14}$$

where P is the number of model parameters.

3 Results

From specified initial values, we first iterate the sampling procedure to burn-in phase of 5000 interactions, already considered thinning of 10 iterations, and the 5000 remaining were used in the analysis. The acceptance rate for update moves, in the Metropolis-Hastings step, is kept around 0.3, this rate was choose based on the convergence of the algorithm.

Figure 1: Scatter plot with marginal histograms of the data.

For the MHDI data set, we observe in Figure 1 that still there are some evidence of heterogeneity in the data. Then, we assume, initially, that the data can be model by a mixture of simplex distribution with two components, that is, we assume $k = 2$.

	Model			
Criteria	M_0	M_1	M_2	M_3
DIC	-6521.251	-11122.33	-6537.52	-11146.01
EAIC	-6505.701	-11112.94	-6523.045	-11133.90
EBIC	-6447.707	-11066.55	-6465.052	-11075.91
Log marginal likelihood	3627.944	5544.9343	3769.03	5553.879

Table 1: Table

Table 1 show the criteria used to compare the four models, including the models without covariate (M_0). In this Table we can observe that the best model, according to these criteria, is the model M_3 with covariate in the mean and dispersion. However, taking into account the parsimonious

criteria we can observe that the model M_1 is more plausible then M_3. Then we choose the model M_1 as a best model among those compared.

Table 2: Number of observations classified across the models and components.

Model		M_0		M_1	
	Component	1	2	1	2
M_0	1	1703	0	1703	0
	2	0	736	622	114
M_1	1	1703	622	2325	0
	2	0	114	0	114

We build the contingency Table 2 to show the classifications of observations in the model M_0 and M_1 and across these models. In addition, the classification of observations can be seen in the scatterplot presented in Figure 2. In the contingency table and in the scatter plot, we can observe that the second component, the component with less weight, decrease in number of observations if the covariates are included in the model. This fact is because more information about the poverty is included in the model changing the distribution of the weights. The distribution of the weights, or probability of the mixture, to model M_0 and M_1 can be seen in the Table 3.

Figure 2: Scatter plot of the classified data.

Finally, table 3 show the posterior mean of the parameters of the modes M_0 and M_1 and 95% HPD credible intervals (Martin *et al.*, 2011). We can observe that the zero is out of range of the HPD interval to β_{11} and β_{12} given evidence that the covariate is significant in the model M_1. The empirical standard deviation is also presented in Table 3 where we can observe that its values are all close to zero given evidence that the all of the parameters are well estimated.

Table 3: Posterior mean, credibility intervals and standard empirical deviation of the estimated parameters.

Model	Parameter	β_{01}	β_{02}	β_{11}	β_{12}	δ_{01}	δ_{02}	δ_{11}	δ_{12}	ω_1	ω_2
M_0	Mean	0.585	0.732	-	-	-2.336	-1.603	-	-	0.693	0.307
	Lower limit	0.583	0.728	-	-	-2.418	-1.740	-	-	0.672	0.287
	Uper limit	0.587	0.735	-	-	-2.255	-1.471	-	-	0.713	0.328
	SD	0.001	0.002	-	-	0.004	0.014	-	-	0.011	0.011
M_1	Mean	1.301	1.598	-1.396	-1.827	-3.045	-3.189	-	-	0.810	0.190
	Lower limit	1.279	1.537	-1.426	-1.924	-3.092	-3.473	-	-	0.714	0.123
	Uper limit	1.318	1.653	-1.360	-1.734	-2.993	-2.948	-	-	0.877	0.286
	SD	0.011	0.031	0.019	0.050	0.027	0.132	-	-	0.043	0.043

References

Chib, S. & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *J. Amer. Statist. Assoc.*, **96**, 270–281.

Fundação Instituto Brasileiro de Geografia e Estatística, D. d. E. e. R. (2014). *Pesquisa nacional por amostra de domicílios, PNAD.: Síntese de indicadores*. IBGE.

Martin, A. D., Quinn, K. M. & Park, J. H. (2011). MCMCpack: Markov chain Monte carlo in R. *Journal of Statistical Software*, **42**(9), 22.

Paz, R. F., Brazan, J. L. & Elher, R. (2014). *A Weibull Mixture Model for the Votes of a Brazilian Political Party*, volume 118 of *Springer Proceedings in Mathematics and Statistics*. Springer.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. pages 583–639.