BETA REGRESSION MODELS: JOINT MEAN AND VARIANCE MODELING

EDILBERTO CEPEDA-CUERVO

Departamento de Estadística [1]

Universidad Nacional de Colombia

Summary

In this paper joint mean and variance beta regression models are proposed. The proposed models are fitted applying Bayesian methodology and assuming normal prior distribution for the regression parameters. An analysis of structural and real data is included, assuming the proposed model, together with a comparison of the result obtained assuming joint modeling of the mean and precision parameters.

Key words: Beta regression, Bayesian methodology, mean and variance modeling

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

1 Introduction

In this paper, we analyze situations where the observations are associated with the beta distribution. The beta distribution defined in equation (1), has applications in uncertainty or random variation of a probability, fraction or prevalence, among others. Thus, this distribution has many applications in areas such as financial sciences or social sciences as education, where random variables are continuous in a bounded interval which is isomorphic to the interval $[0, 1]$. To mention an example, in studies of the quality of education, a number from 0 to 5 (or any other positive integer bounds) is assigned as a measure of performance for the evaluation of school subjects as math, language, arts, natural sciences or any other scholar area. In these cases, the measure assigned to each student can be expressed as a number from zero to one. Thus, it can be assumed that the level of student performance is a random variable with beta distribution.

The beta p, q distribution function, defined by equation (1) can be re-parametrized as a function of the mean and the so called dispersion parameter as in equation (4), or as function of the mean and variance taking into account equations (5) and (6). This characterization of the beta distribution can be more appropriate. In the first re-parametrization, making $\phi = p + q$ we may 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 presented in Ferrari and Cribari-Neto (2004), was already proposed in the literature, for example in Jorgensen (1997) or in Cepeda (2001, pg 63).

In this case, the mean and dispersion parameters can be modeled as functions of explanatory variables, given that behavior of these parameters can be explained explanatory variables. To cite a few examples, the educational level of mothers could influence students school performance; land concentration can be explained by random variables associated with social and political factors or the proportion of income spent monthly could be explained by the number of persons in the household. At the same time, we can assume that the dispersion parameter changes as a function of the same or other random variables. With these ideas, Bayesian regression, with joint modeling of the mean and dispersion parameters, was initially proposed by Cepeda (2001, pg. 63), under the framework of joint modeling in the biparametric exponential family (see Cepeda and Gamerman 2001, 2005). After that, Ferrari and Cribari-Neto (2004) proposed classical beta regression models, assuming that the dispersion parameter is constant through the rank of the explanatory variables. Further works have been published by Smithson and Verkuilen (2006), Simas et al. (2010) and, Cepeda-Cuervo and Achcar (2010), the latter proposing nonlinear beta regression in the context of Double Generalized Nonlinear Models. The beta regression models were extended in Cepeda et al.(2011), assuming that the observation are spatially correlated.

The rest of the paper is organized as follows: Section 2 includes general concepts on beta distribution. Section 3, presents the joint mean and variance beta regression models. Section 4, provides an analysis of the structural data assuming nonlinear and logistic regression models. Section 5, presents the results of the "language performance" data.

2 Beta Distribution

A random variable Y has 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. The mean and variance of Y, $\mu = E(Y)$ and $\sigma^2 = Var(Y)$, are given by

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

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

Many random variables can be assumed to have beta distribution. For example, income inequality or land distribution when measured using the Gini index proposed by Atkinson(1970), and the performance of students in subjects such as mathematics, natural sciences or literature. In the latter case, if performance X takes values within the interval (a, b), the random variable $Y = (X - a)/(b - a)$ can be assumed to have beta distribution. This performance can be explained by household socioeconomic variables, having fundamental impact on the student cognitive achievement. For example, the level of student achievement is closely related to the educational level of their parents and the number of hours devoted to study a subject. Thus, the beta regression model could be appropriate to explain the behavior of school performance as a function of associated factors. In these applications however, the reparametrization of the beta distribution given in (4) could be more appropriate. In the first, doing $\phi = p + q$ we can see that $p = \mu\phi$, $q = \phi(1 - \mu)$ and $\sigma^2 = \frac{\mu(1-\mu)}{\phi+1}$. Hence, 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 reparametrizacion that is presented in Ferrari and Cribari-Neto (2004), had already appeared 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

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

In this case, the mean and dispersion parameters can be modeled as function of explanatory variables, for example, as was proposed in Cepeda(2001), given that changes in the precision parameter can be explained by explanatory variables, such as mothers educational level in the case of the student's school performance.

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

$$p = \frac{(1-\mu)\mu^2 - \mu\sigma^2}{\sigma^2} \tag{5}$$

$$q = \frac{(1-\mu)[\mu - \mu^2 - \sigma^2]}{\sigma^2} \tag{6}$$

Although writing (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 achieved applying the Bayesian methodology proposed in Cepeda(2001), and Cepeda and Gamerman (2005). Sometimes, joint modeling of the mean and variance could be more appropriate than the joint modeling of the mean and the so

called dispersion parameter, given that parameters of the regression models would be more easily interpreted.

3 Joint Mean and Variance Beta Regression Models

With the reparametrization of the beta distribution as a function of μ and ϕ, we can define a double generalized beta regression model as proposed in Cepeda (2001). In that research, joint modeling of the mean and dispersion parameters in the beta regression model and a Bayesian methodology to fit the parameters of the proposed model, was defined. Under a general framework, a random sample $Y_i \sim Beta(p_i, q_i)$, $i = 1, 2, \ldots, n$, was assumed, where both, mean and precision parameters, are modeled as a function of explanatory variables. That is,

$$\text{logit}(\mu) = \mathbf{x}_i^t \boldsymbol{\beta} \tag{7}$$

$$\log(\phi) = \mathbf{z}_i^t \boldsymbol{\gamma} \tag{8}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_p)$ are the vectors of the mean and dispersion regression models and, $\mathbf{x}_i$ and $\mathbf{z}_i$ are the vectors of the mean and dispersion explanatory variables, at the i-th observation, respectively. Afther Cepeda's work, Ferrari and Cribari-Neto (2004) proposed the same reparametrization of the beta distribution, $\mu = p/(p+q)$ and $\phi = p + q$. In that paper, they assumed that $g(\mu_i) = \mathbf{x}_i^t \boldsymbol{\beta}$, where g is a strictly monotonic and twice differentiable real valued link function defined in the interval $(0, 1)$, assuming that the dispersion parameter is constant. Although they consider

many possible link functions, in the applications they take the logit link function, given that the mean can be interpreted as a function of the odds ratio. The joint mean and dispersion beta regression models proposed by Cepeda(2001), was later studied by Smithson and Verkuilen (2006) and Simas et al. (2010), under a classical perspective. At the same time, a nonlinear beta regression was proposed by Cepeda and Achcar (2010), assuming a nonlinear mean model given by (9) and a dispersion model given by (8), in the context of Double Generalized Nonlinear Models. This model was applied to the schooling rate data analysis in Colombia, for the period ranging from 1991 to 2003.

$$\mu_i = \frac{\beta_0}{1+\beta_1 \exp(\beta_2 x_i)} \tag{9}$$

In this paper, we propose joint mean and variance beta regression models, with the mean modeled as linear or nonlinear function of the parameters, as in (7) or (9), and the variance modeled as a function of the explanatory variables (10), where g is a monotonic and two time differentiable real function, that take into account the positivity of the variance.

$$g(\sigma_i^2) = \mathbf{z_i}^t \boldsymbol{\gamma} \tag{10}$$

The results of fitting the mean and variance beta regression models are easily interpretable: the mean fitted models have the usual interpretation, but the fitted variance model is easily interpreted directly from data behavior. For example, if the explanatory variable Z_1 is associated to γ_1 and $\gamma_1 > 0$, increasing behavior of Z_1 is associated with increasing behavior of σ^2. In

the same way, the interpretation is applicable when the parameters of the variance models are negative.

In the next sections, structured and real data sets are analyzed applying joint mean and dispersion, and joint mean and dispersion beta regression models to compare the performance of these models, according to the behavior of the data.

4 Structural Data Analysis

In this section we present the results of the studies of a structural data set. The aim is to fit joint nonlinear (logistic) mean and variance regression models and compare the results with the results obtained when joint nonlinear (logistic) mean and dispersion models are fitted to the same data.

The data set, represented by black points in Figure 1, were generated assuming as explanatory variable X that takes values from 1 to 13. Interest variable Y, that increases with X, is assumed to have beta distribution. Through X, Y it presents an increase variance.

4.1 Beta Nonlinear Regression

4.1.1 Joint nonlinear mean and variance beta regression models

In this section we assume that the observations come from the beta distribution. Exactly, we assume that $Y_i \sim Beta(p, q)$, $i = 1, 2, \ldots, n$, where $\mu_i = E(Y_i)$ and $\sigma_i^2 = \mathrm{Var}(Y_i)$, follow the models given by (9) and $\log(\sigma_i^2) = \gamma_0 + \gamma_1 x_i$, respectively. Assuming independent normal prior distribution for the regression parameters, 5.000 samples of the posterior distribution were

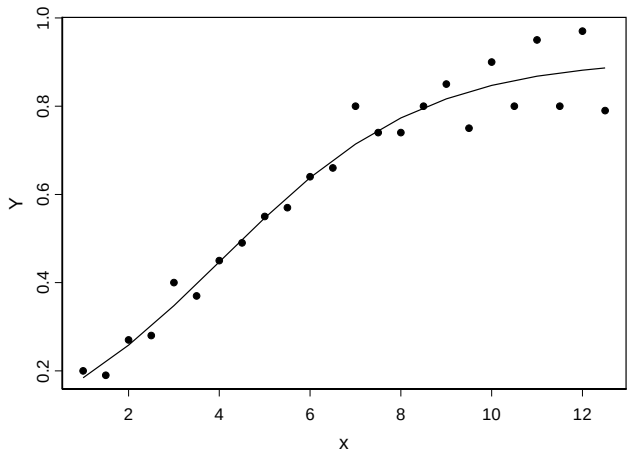

Figure 1: Systematic data (black points) and posterior fit mean model (continuous line), given by (9) .

generated, using WinBugs software, Spiegelhalter et al., (2002). The posterior parameter estimates were obtained from the sample of the posterior distribution taking an observation each five, after a burning of 1.000 observations. The posterior parameter estimates given by the mean of the posterior samples are given in Table (1). For this model, the logarithm of the likelihood function is given by $2logL = -346.230$, and the value of the Deviance Information Criterion (DIC) is equal to -336.916.

Figure 1, shows good agreement between data and the fit mean model. The variance takes small values that increase with X, given that estimation of γ_1 is positive, following the general behavior of the data.

Parameters	Mean model			Variance model	
	β_0	β_1	β_2	γ_0	γ_1
mean	0.9073	6.12	-0.4456	-8.078	0.2572
s.d.	0.0171	0.2658	0.0169	0.3098	0.0408

Table 1: Parameter estimates of mean and variance regression parameters

4.1.2 Joint Nonlinear Mean and Precision Beta Regression Models

In this section, we assume that interest variable data comes from beta distribution $Y_i \sim Beta(p, q)$, $i = 1, 2, \ldots, n$, where the mean model is given by (9) and the dispersion model by $\log(\phi_i) = \gamma_0 + \gamma_1 x_i$, for the purpose of comparing variation in the posterior Bayesian summaries, obtained when nonlinear beta regression models with joint modeling of the mean and dispersion parameters, are fitted with results obtained in Section 4.1.1.

For this model, the posterior parameter estimates and the respective standard deviation, obtained by proceeding as in Section 4.1.1, and assuming the same normal prior distribution function, are given in Table 2. In this case, the $2logL = -340.602$ and the DIC criterion value is equal to -331.080.

4.1.3 Model comparison

From Sections 4.1 and 4.2, it is possible to conclude that the beta nonlinear regression model, with joint modeling of the mean and variance, has greater likelihood value and smaller DIC value than the beta nonlinear regression model with joint modeling of the mean and precision parameters. Thus,

Parameters	Mean model			Precision model	
	β_0	β_1	β_2	γ_0	γ_1
mean	0.9058	6.1120	-0.4463	6.7060	0.2876
s.d.	0.0172	0.2414	0.0168	-0.3259	0.0368

Table 2: Parameter estimates of joint mean and variance parameters

between these models, the first one is better to fit the proposed structural data set.

Figure 2, shows the behavior of variance as per joint mean and variance modeling (continuous line) and the joint mean and precision models (dashed line). Although in both cases variance increases with X, the general behavior disagrees, given that when the variance is directly modeling the variance of data behavior is better described, especially for smaller and bigger values of X. However, the fitted mean models present smaller differences.

4.2 Beta Logistic Regression Models

In this section, we analyze the systematic data set applying the proposed beta regression models, assuming joint modeling of the mean and variance parameters, and the beta regression models assuming joint modeling of the mean and dispersion parameters, but with logistic mean models in both cases. From the posterior estimates of the parameters, the performance of the models are compared to determine which model fits the data set better.

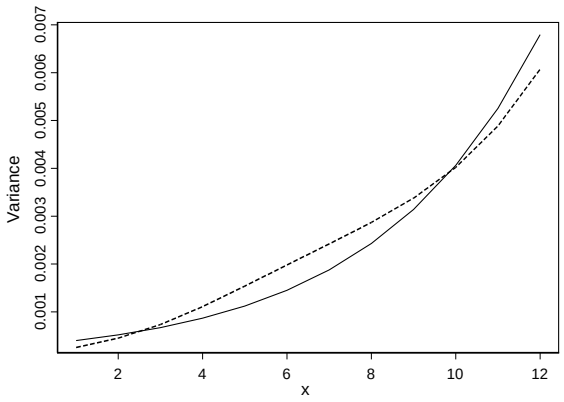

Figure 2: Variance models comparison: variance from the joint mean and variance model (continuous line) and variance from mean and precision models (dotted line)

4.2.1 Joint Mean and Variance Beta Regression Models

In this section, we assume that the interest variable follows beta distribution $Y_i \sim Beta(p, q)$, $i = 1, 2, \ldots, n$, where the mean and variance models are given by (11) and (12), respectively.

$$logit(\mu_i) = \beta_0 + \beta_1 x_i \quad \text{and} \tag{11}$$

$$\log(\sigma_i^2) = \gamma_0 + \gamma_1 x_i \tag{12}$$

The posterior mean of the parameter samples and the respective standard deviation are given in Table 3. For this model, $2logL = -320.016$ and the DIC criterion value is equal to -311.922. The fit mean and variance model given by (11) and (12) are represented by continuous line in Figures 3 and 4.

Parameters	Mean model		Variance model	
	β_0	β_1	γ_0	γ_1
mean	-1.761	0.3764	-7.605	0.1847
s.d.	0.0417	0.0093	0.3059	0.0355

Table 3: Parameter estimates of joint mean and variance parameters for beta regression models (11) and (12).

4.2.2 Joint Mean and Precision Beta Regression Models

In this section, we assume that the interest variable data comes from the beta distribution $Y_i \sim Beta(p, q)$, $i = 1, 2, \ldots, n$, where the mean model is given by (11) and the precision by $\log(\phi_i) = \gamma_0 + \gamma_1 x_i$. This, for the purpose of comparing the posterior Bayesian summaries obtained fitting joint mean and precision models with the posterior summaries obtained in Section 4.2.1, where joint mean and variance beta regression models were fitted. The posterior inferences of the parameters were obtained as in the lather sections, assuming the same independent normal prior distribution, and are given in Table 4. For this model $-2logL = 313.442$ and the DIC value criterion is equal to $DIC = -305.304$. The fit mean and variance obtained from (11) and $\log(\phi_i) = \gamma_0 + \gamma_1 x_i$, are represented by dotted line in figures 3 and 4.

4.2.3 Models Comparison

Between the models fitted in Sections (4.2.1) and (4.2.2) it is possible to conclude that the beta logistic regression model, with joint modeling of the mean and variance, has greater likelihood value and smaller DIC value than

Parameters	Mean model		Precision model	
	β_0	β_1	γ_0	γ_1
mean	-1.779	0.3762	6.502	-0.3151
s.d.	0.0413	0.0101	0.3074	0.0374

Table 4: Posterior parameter estimates of joint mean and precision parameters.

beta logistic regression model, with joint modeling of the mean and precision parameters. Thus, this model is the one that best fit the proposed structural data set.

Figure 3, shows the behavior of the fitted mean models for the joint mean and variance modeling (continuous line) and for the joint mean and precision models (dashed line). From this figure, it is clear that the joint mean and variance model is the model that best fits this structural data set. This conclusion may also be drawn from Figure (4), where the continuous line is a better description of the variance data behavior. Although in both cases the variance increases with X, the general behavior of the dotted line disagrees, showing that, when the variance is directly modeled the variance of data behavior is better described, particularly for smaller values of X.

In each of the cases considered in this study, several chains were generated starting from different initial values. All of them provided a rough indication of convergence after a small transient period. Although, the joint mean and variance regression models proved be more sensitive to initial values, these models can be seen more appropriately in this data analysis. In general,

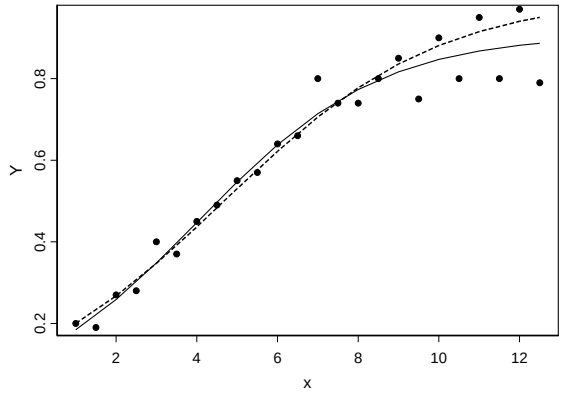

Figure 3: Fit mean models given by 3. Mean and variance model (continuous line). Mean and precision model (dotted line).

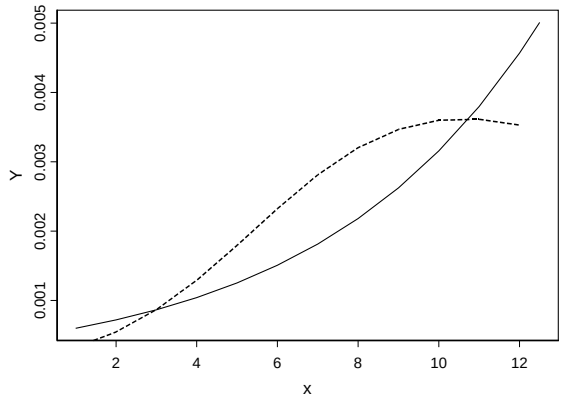

Figure 4: Variance models comparison for logistic models: Joint mean and variance modeling (continuous line) and Joint logistic and Precision models (dotted line)

these models can be formulated from a descriptive analysis of the data set. For example, from the plot of this data set it is easy to conclude that the mean follow a non nonlinear model and that the variance are increases with X. Other usual behaviors may also be easily determined. For example, if the variance decreases with X or if it increases to a real value c, after which it is decreases. Thus, the joint mean and variance models should be taking into account when data sets are analyzed applying nonlinear regression beta models.

5 Application

In this section, we present the results of the analysis of a data set which consists of the mean performance in Spanish of students in 31 departments of Colombia, obtained from the Ministry of Education (MEN) and from National Institute of Statistics (DANE), calculated from the National Household and Population Census in 2005. The interest variable is the mean performance "Performance" in Spanish of students in second grade of secondary schools, and the explanatory variables are the level of unsatisfied basic needs UBN and the percentage of teachers with postgraduate levels of educations.

The data behavior is presented in Figure 5. The first, shows that the Spanish performance is a decreasing function of UNB and that the variance is constant through UNB. The second, shows that performance is an increasing function of $PERC$ and that variance change with $PERC$, in increasing manner.

Although we initially assumed joint mean and variance (dispersion) mod-

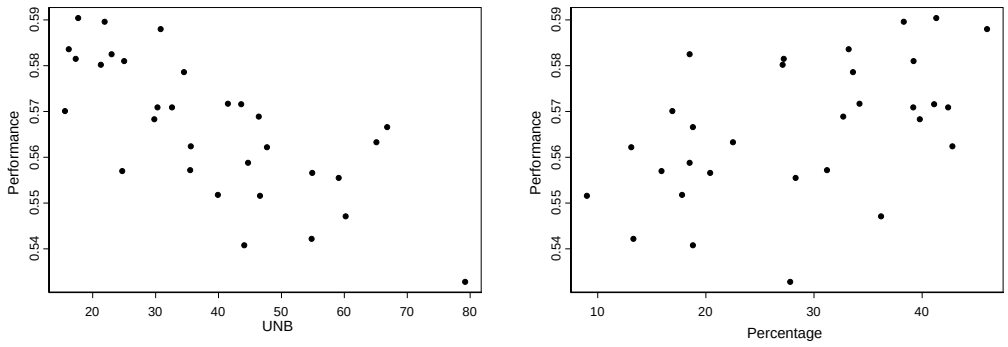

Figure 5: Plots of performance in Spanish versus explanatory variables

eling, including all explanatory variables, we present the result of the beta regression model with mean and variance models given by equations (13) and (14), respectively, given that the DIC value of the second models was smaller than the one for the first models.

$$\text{logit}(\mu) = \beta_0 + \beta_1 NBI + \beta_2 PER \quad (13)$$

$$\log(\sigma^2) = \gamma_0 + \gamma_2 PER \quad (14)$$

Assuming normal prior distribution $\beta_i \sim N(0, 10^2)$, $i = 0, 1, 2$ and $\gamma_i \sim N(0, 10^2)$, $i = 0, 1$, for the parameters, 10.000 samples of the posterior distribution were generated. The parameter estimates were obtained from the posterior sample, after burning off the first of 1.000 samples. Parameter estimates and the corresponding standard deviations are given in Table 5. For this model, $2logL = 205.323$ and the DIC value is equal to -195.769. When a beta regression model without explanatory variables in the variance model is assumed, $2logL = 204.144$, the DIC value is equal to -196.222.

DIC	Parameters	Mean model			Variance model	
		β_0	β_1	β_2	γ_0	γ_1
-196.222	$\hat{\theta}$	0.3132	-0.0025	0.0018	-8.425	-0.0306
	s.d.	0.0357	5.023E-4	7.564E-4	0.8152	0.0269
-195.769	$\hat{\theta}$	0.3026	-0.0023	0.0019	-9.287	-
	s.s.	0.0316	607E-4	6.952E-4	0.2766	-

Table 5: Estimates of the parameters of the variance models

Table 5, includes the estimates of the mean and dispersion models given by equations (13) and $\log(\sigma^2) = \gamma_0 + \gamma_2 NUM$, respectively. With the same prior distribution, the posterior parameter estimates obtained in this case are given in Table . For this model, $2logL = 205.358$ and the DIC value is equal to -195.464. We also considered the model without explanatory variables in the precision model, for which $2logL = 204.212$ and the DIC value is equal to -196.186

DIC	Parameters	Mean model			Precision model	
		β_0	β_1	β_2	γ_0	γ_1
-195.464	$\hat{\theta}$	0.315	-0.0025	0.0017	7.022	0.0289
	s.d.	0.0369	5.091E-4	7.686E-4	0.7897	0.0265
-196.186	$\hat{\theta}$	0.3064	-0.0023	0.0018	7.881	-
	s.d.	0.03172	4.73E-4	6.921E-4	0.2601	-

Table 6: Estimates of the parameters of the precision models

This application shows how the joint mean and variance beta regression models can be proposed easily from the data behavior. Shows also that the proposed model fit the data better than the joint mean and precision models. This result show the performance of the proposed models in the analyze this type of data set.

References

[1] Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*,**2**, 244-263.

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

[3] 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.

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

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

[6] Cepeda-cuervo, E., Nuñez-Anton V. (2011). Generalized Econometric Spatial Models. Commun. Stat., Simulation Comput. 39, No. 2, 405-419.

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

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

[9] 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.

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

[11] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, **64**4, 583-639.