مجله‌ی پژوهش‌های آماری ایران
پژوهشکده‌ی آمار دوره‌ی ۱۴، شماره‌ی ۱، بهار و تابستان ۱۳۹۶، صص ۱۰۱–۱۱۸

J. Statist. Res. Iran **14** (2017): 101–118
DOI: 10.18869/acadpub.jsri.14.1.101

Bayesian Analysis of Augmented Mixed Beta Models with Skew-Normal Random Effects

Zohreh Fallah Mohsenkhani[†,‡,*], Mohsen Mohammadzadeh[†] and Taban Baghfalaki[†]

[†] Tarbiat Modares University
[‡] Statistical Research and Training Center

Received: 11/6/2016 Approved: 6/4/2017

Abstract. Many studies in different areas include data in the form of rates or proportions that should be analyzed. The data may also accept values zero and one. Augmented beta regression models are an appropriate choice for continuous response variables in the closed unit interval $[0, 1]$. The data in this model are based on a combination of three distributions, degenerate distribution at 0 and 1, and a beta density in $(0, 1)$. The random effects are usually added to the model for accommodating the data structures as well as correlation impacts. In most of these models, the random effects are generally assumed to be normally distributed, while this assumption is frequently violated in applied studies. In this paper, the augmented mixed beta regression model with skew-normal distributed random effects is presented. A Bayesian approach is adopted for parameter estimation using Markov Chain Monte Carlo method. The proposed model is applied to analyze a real data set from Labor Force Survey.

Keywords. Augmented beta regression; beta distribution; mixed models; Bayesian approach; skew-normal distribution.

MSC 2010: 62J12; 62P20.

[*] Corresponding author

1 Introduction

Beta regression models are commonly used to model variables that assume values in the open interval $(0, 1)$. They are based on the assumption that the dependent variable follows a beta distribution, also its mean is related to a set of explanatory variables through a linear predictor with unknown coefficients and a suitable link function, (Figueroa-Z et al., 2013). The beta distribution is flexible for modeling proportions and rates since its density has different shapes depending on the values of the two parameters. Many studies have been done for observed outcomes containing zero and one which are not in the support of the beta distribution. Inflated beta distributions were explained as a mixed continuous-discrete distribution to model proportion values that contain zeros and ones by Ospina and Ferrari (2010). Inflation happens when the probability mass of some points exceeds that of a proposed distribution. For example, zero-inflated counts or zero inflated binomial regression models that were introduced by Lachenbruch (2002) and Hall and Berenhaut (2002), respectively. In beta distribution which is a continues one, the model with extra zeros and ones is preferred to be called as augmented points instead of inflated points at zero and one (Galvis et al., 2014). They were explained augmented beta regression by mixed continuous-discrete distribution to model proportion values containing zeros and ones. Verkuilen and Smithson (2012) extended generalized linear mixed models for beta distributed response variables. Mixed beta regression model under the Bayesian approach while including fixed and random effects are discussed by (Figueroa-Z et al., 2013). Random effects models are commonly used to model correlated data. In longitudinal data, these effects are usually assumed to be normally distributed. Some studies showed that the normality assumption is inappropriate when data exhibit non-normal behavior (Zhang and Davidian, 2001). Thus, some extension of statistical models with remarkable flexibility in the distributional assumptions are required that can readily adapt to the non normality behavior of the random effects. Azzalini (1985) presented a flexible family of skew-normal distributions that contains the normal family as a special case. Studies on the multivariate skew-normal distributions are also presented by (Azzalini and Dalla Valle, 1996). The skew-normal linear mixed models are presented by Arellano-Valle et al. (2005), Lachos et al. (2010) and Lin (2008).

The augmented mixed beta models was defined under normality assumption for random effects which is an extension of linear mixed effect model.

A commonly used approach for analyzing proportion cluster data such as longitudinal proportion is augmented mixed beta models. In this paper, we have propose using a skew-normal random effects instead of the normal one to allow more flexibility in the augmented mixed beta regression models. So we consider a zero and one augmented Beta regression model with a skew-normal distributional assumption for random effects. This model is amenable to the Bayesian implementation and the MCMC scheme described for sampling from the posterior distribution. We have used the available software WinBUGS and R package for fitting the models, where the criteria of Log Pseudo Marginal Likelihood (LPML) have been used for model comparison.

The motivating data example in this paper comes from a Labor Force Survey (LFS) data to determine the proportion of employed persons in the household. For each household, this variable is measured four times and since these data are considered as longitudinal data. We need to investigate random effects in the model for considering the correlation among the response variable during the time. In the application Section, we intend to compare the performance of the augmented mixed beta regression model under normal and skew-normal random effects and we conclude that the augmented mixed beta regression under skew-normal random effects is a better fit compared to other one.

The rest of the paper proceeds as follows. After a brief introduction to the beta regression model and the augmented mixed beta regression model under normal distribution for random effects, Section 2 introduces the augmented mixed Beta models with skew-normal random effects. Section 3 develops the Bayesian estimation scheme. In Section 4 some Bayesian criteria have been presented for model selection. In Section 5, some simulation studies are carried out to examine the performance of the proposed model. Section 6 contains an application using an Iranian LFS data set and concluding remarks are given in Section 7.

2 Statistical Model and Bayesian Inference

In this section, the beta regression model is introduced for responses lying in $(0, 1)$. Then an augmented mixed beta regression model with normal random effects is presented for modeling the observations in the closed interval $[0, 1]$. Next, we utilize skew-normal distributions for random effects in the augmented mixed beta regression model as a tool for robust modeling under a Bayesian approach.

2.1 Beta Regression Model

The beta distribution is often applied for modeling continuous data that are confined in the interval $(0, 1)$. A random variable Y follows a beta distribution parameterized in terms of its mean and precision parameter ϕ is given by

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \tag{1}$$

where $0 < y < 1$, $0 < \mu < 1$ and $\phi > 0$. If $Y \sim \text{beta}(\mu\phi, (1-\mu)\phi)$, then $E(Y) = \mu$ and $Var(Y) = \mu(1-\mu)/(1+\phi)$. A beta regression model can be considered under a generalized linear model (GLM) framework by linking the subject-specific mean μ_i and covariate vector $\mathbf{X}_i$ (Cribari-Neto and Zeilis, 2010). This is given as $g(\mu_i) = \mathbf{X}_i^T\beta$, where β is the vector of regression parameters.

2.2 Augmented Mixed Beta Regression Models

In some studies the data may lie in $[0, 1]$. For modeling such data the augmented beta regression is defined by (Galvis et al., 2014) that comprises of a mixture of three distributions, with two degenerated distributions at 0 and 1, and a beta distribution. Thus the density of $Y_i, i = 1, \ldots, n$, is given by:

$$f(y_i; p_0, p_1, \mu_i, \phi) = \begin{cases} p_0 & y_i = 0 \\ p_1 & y_i = 1 \\ (1-p_0-p_1) f(y_i; \mu_i, \phi) & y_i \in (0,1), \end{cases} \tag{2}$$

where $p_0 = P(Y_i = 0)$, $p_1 = P(Y_i = 1)$ and $f(y_i; \mu_i, \phi)$ is given in (1). We denote the density of Y_i by $AB(p_0, p_1, \mu\phi, (1-\mu)\phi)$ which using indicator functions would be written as follows:

$$f(y_i; p_0, p_1, \mu_i, \phi) = p_0^{I(y_i=0)} p_1^{I(y_i=1)} \{(1-p_0-p_1) f(y_i; \mu_i, \phi)\}^{(1-I(y_i=1))(1-I(y_i=0))} \tag{3}$$

where $I(\cdot)$ denotes the indicator function. The mean of Y_i and its variance can be derived as:

$$E(Y_i) = (1-p_0-p_1)\mu_i + p_1,$$
$$Var(Y_i) = p_1(1-p_1) + (1-p_0-p_1)[\frac{(1-\mu_i)\mu_i}{(1+\phi)} + (p_0+p_1)\mu_i^2 - 2\mu_i p_1].$$

The augmented mixed beta regression model is defined as follows. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be independent continuous random vectors, where $\mathbf{Y}_i = (y_{i1}, \ldots, y_{in_i})$ represents an observed response vector for a sample unit i, with components, y_{ij} in the interval $[0, 1]$. The regression model is obtained over a suitably transformed $\boldsymbol{\mu_i}$, with the following structure by (Galvis et al., 2014).

$$G(E(\mathbf{Y}_i; \mathbf{b}_i)) = g(\boldsymbol{\mu_i}) = \mathbf{X}_i^T \beta + \mathbf{Z}_i^T \mathbf{b}_i, \qquad i = 1, \ldots, n, \tag{4}$$

where $\mathbf{X}_i$ is the design matrix of dimension $p \times n_i$ corresponding to the vector of fixed effects $\beta = (\beta_1, \ldots, \beta_p)^T$, and $\mathbf{Z}_i$ is the design matrix of dimension $q \times n_i$ corresponding to the vector of random effects $\mathbf{b}_i = (b_{i1}, \ldots, b_{iq})^T$.

2.3 Augmented Mixed Beta Models with Skew-Normal Random Effects

The normality of the random effects is a routine assumption for the GLM framework, but it may be unrealistic. We have utilized skew-normal distribution as a tool for robust modeling under a Bayesian approach. A random vector $\mathbf{b}$ follows a multivariate skew-normal distribution with location vector $\boldsymbol{\mu}$, covariance matrix Σ_b and skewness parameter vector $\boldsymbol{\lambda}$, if its pdf is given by:

$$\pi(\mathbf{b}) = 2\phi_q(\mathbf{b}; \boldsymbol{\mu}, \Sigma_b)\Phi(\boldsymbol{\lambda}^T \Sigma_b^{-1/2}(\mathbf{b} - \boldsymbol{\mu})), \tag{5}$$

where $\phi_p(\cdot; \boldsymbol{\mu}, \Sigma_b)$ is the density function of the q-variate normal distribution and $\Phi(\cdot)$ is the cumulative distribution function of the standard univariate normal distribution (Azzalini and Dalla Valle, 1996). Now consider the augmented mixed beta model with a single random intercept, where the random effects has pdf as in (5) with q=1, that is $b_i \sim SN(0, \sigma_b^2, \lambda)$. A particularly useful and easy interpretable link function for proportional data is the logit link function (Ferrari and Cribari-Neto, 2004), therefore the logit link function is used.

For the logit link function we can consider the following structure:

$$\text{logit}(\boldsymbol{\mu_i}) = \mathbf{X}_i^T \beta + \mathbf{Z}_i^T b_i. \tag{6}$$

3 Model fitting Using MCMC Sampling

We use the Bayesian methodology using MCMC sampling for the Augmented mixed beta regression model with skew-normal distributional assumptions for random effects since a full Bayesian approach is difficult to implement for such models. One MCMC type approach, that needs only the characteristics of the conditional posterior distribution of each parameter, is the Gibbs sampler.

For convenience, we suppose that elements of θ in the model are independent and are defined by $\theta = (\beta, p_0, p_1, \phi, \sigma_b^2, \lambda)$. The primary goal here is to estimate θ. Considering the random effects design matrix $\mathbf{Z}_i$ to be an identity matrix, $(\mathbf{y}_1, \mathbf{x}_1), \ldots, (\mathbf{y}_n, \mathbf{x}_n)$ are the observed sample for n subjects, with $\mathbf{y}_i$ as the response vector for subject i. Considering the fixed effects as β, a Bayesian specification of the model requires consideration of the prior distribution for all parameters. We attempt to assign weakly informative distribution for the parameters. Also, the hyper parameters are selected such that they lead to the low informative prior distributions. We adopt the following prior specifications:

$$\begin{aligned} \beta &\sim N_p(0, \Sigma_\beta), & \Sigma_\beta &= diag(100, 100, 100) \\ p_0 &\sim U(0, 1), & p_1 &\sim U(0, 1 - p_0) \\ \phi &\sim \Gamma(0.01, 0.01), & \sigma_b^2 &\sim I\Gamma(0.01, 0.01) \\ \lambda &\sim HN(0, 100), \end{aligned} \tag{7}$$

where HN denotes the half-Normal distribution, and $I\Gamma$ denotes the inverse gamma distribution. Let $b = (b_1, b_2, \ldots, b_n)$, the complete conditional likelihood of the outcome y is given by:

$$\begin{aligned} L(\theta|y, X, b) &= \prod_{i=1}^{n} \prod_{j=1}^{n_i} f(y_{ij}|b_i) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{n_i} p_0^{I(y_{ij}=0)} p_1^{I(y_{ij}=1)} \{(1 - p_0 - p_1) f(y_{ij}; \mu_{ij}, \phi)\}^{(1-I(y_{ij}=0))(1-I(y_{ij}=1))} \end{aligned}$$

where $\theta = (\beta, p_0, p_1, \phi, \sigma_b^2, \lambda)$ denotes the unknown model parameters vector. For convenience, we use the stochastic representation of the skew-normal

distribution in which

$$b \stackrel{def}{=} \lambda|Z_0| + Z_1, \tag{8}$$

where $Z_0 \sim N(0,1), Z_1 \sim N(0,\sigma_b^2)$ and $\stackrel{def}{=}$ meaning 'distributed as', with independent Z_0 and Z_1 by Arellano-Valle et al. (2005).

The joint posterior distribution using the Bayes Theorem is given by

$$\pi(\theta, b|y, X) \propto L(\theta|y, X, b)g(b; z_0, \sigma_b^2)\pi(\beta)\pi(p_0)\pi(p_1)\pi(\phi)\pi(z_0)\pi(\sigma_b^2), \tag{9}$$

where

$$g(b; z_0, \sigma_b^2) = g(b; z_0)\pi(z_0; \sigma_b^2)\pi(\sigma_b^2).$$

The posterior distribution (9) is analytically intractable, but the MCMC methods such as the Gibbs sampler and Metropolis-Hastings algorithm can be used to draw samples. The Gibbs sampler works by drawing samples iteratively through full conditional distributions derived from (9). Since some of the full conditional distributions are not available, the Metropolis–Hastings algorithm is applied. We implement the Bayesian methodology using MCMC techniques for the beta Regression model with the skew-normal distributional assumption for random effects. A key feature of this model, which allows writing WinBugs codes, is that it can be formulated in a flexible hierarchical representation as follows:

$$\begin{aligned}
&Y_{ij}|\mu_{ij} \sim AB(p_0, p_1, \mu_{ij}\phi, (1-\mu_{ij})\phi),\\
&\text{logit}(\mu_{ij}) = \mathbf{X}_i^T\beta + b_i,\\
&[b_i|z_0, \sigma_b^2] \sim N(\lambda z_0, \sigma_b^2),\\
&Z_0 \sim HN(0,1).
\end{aligned}$$

The relevant MCMC steps were implemented using the R2WinBUGS package which connects the R with the WinBUGS software.

4 Bayesian Model Selection

Many criteria have been defined for model selection in the Bayesian inference. One of the popular criterion, which is usually used is the conditional predictive ordinate (CPO) statistic. Let the vector Y be the data set, $Y^{(-i)}, i = 1, \cdots, n$ denote all data except the ith observation and $\pi(\Theta|Y^{(-i)})$

denotes the posterior distribution of the model parameter vector Θ given $Y^{(-i)}$. For each i, the criterion $\text{CPO}_i = \int_\Theta f(y_i|\theta)\pi(\theta|Y^{(-i)})d\theta$ can be obtained by $\text{CPO}_i = (\frac{1}{m}\sum_{k=1}^{m}\frac{1}{f(y_i|\theta^{(k)},Y)})^{-1}$, where m is the number of iterations after a burn-in period (Gelfand and Dey, 1994). A statistic of CPO_i is the LPML, defined by $\text{LPML} = \sum_{i=1}^{n}\log(\text{CPO}_i)$. The larger values of LPML indicates the better fit for the model. Two other criteria EAIC and EBIC (Carlin and Louis, 2008) can also be used for model selection. Let $\theta^{(k)}$ be the MCMC posterior sample generated at the iteration k of the algorithm, η is the number of parameters, n is the number of observations and $\overline{D} = \frac{1}{m}\sum_{k=1}^{m}\sum_{i=1}^{n}\log f(y_i|\theta^{(k)},Y)$. Then $EAIC = -2\overline{D} + 2\eta$, and $EBIC = -2\overline{D} + 2\eta\log n$. Unlike LPML, the smaller values of EAIC and EBIC indicate a better fit for the model.

In order to assess the convergence of MCMC algorithm, we have used the Gelman and Rubin's diagnostic test, (Gelman and Rubin, 1992). This method requires two or more parallel chains (the number of which is denoted by m') to be generated, each with different starting values. For assessing convergence of individual model parameters, θ, this convergence diagnostic test is proposed as a univariate statistic, referred to as the potential scale reduction factor (PSRF). Calculation of this ststistic is based on the last n' samples in each of m' parallel chains. This factor is calculated by:

$$PSRF = \left(\frac{n'-1}{n'} + \frac{m'+1}{n'm'}\cdot\frac{B}{W}\right)^{\frac{1}{2}},$$

where (B/n') is the between-chain variance and W is the within-chain variance and are calculated by:

$$B = \frac{n'}{m'-1}\sum_{j=1}^{m'}(\bar{\theta}_{\cdot j} - \bar{\theta}_{\cdot\cdot})^2,$$

$$W = \frac{1}{m'}\sum_{j=1}^{m'}\left[\frac{1}{n'-1}\sum_{i=1}^{n'}(\theta_{ij} - \bar{\theta}_{\cdot j})^2\right].$$

As chains converge to a target distribution, the estimated variance $\hat{V} = (1-\frac{1}{n'})W + \frac{1}{n'}B$ should be almost equivalent to the within-chain variance, so PSRF should approximately equal one.

5 Simulation Study

Some simulation studies are illustrated to assess the performance of the proposed models. In all cases, we generate 100 data sets with sample size $n = 100$ with $n_i = 5$, $i = 1, \ldots, 100$. To generate the dependent variable with $p_0 = p_1 = 0.1$, first we draw a Bernoulli sequence with success probability $(1 - p_0 - p_1)$ that is allocated to $y_{ij} \in (0, 1)$. Then, conditional on this draw, we carry out another Bernoulli with equal success probability for one-occurences. In this way, the sequence of $y_{ij} \in (0, 1)$, $y_{ij} = 0$ and $y_{ij} = 1$ with $p_0 = 0.1$, $p_1 = 0.1$ and $(1 - p_0 - p_1) = 0.8$ are generated. We took the augmented mixed beta regression model with location parameter μ_{ij}. For this purpose, μ_{ij} is generated as: $\text{logit}(\mu_{ij}) = \beta_1 + \beta_2 x_{2ij} + \beta_3 x_{3ij} + b_i$ where $i = 1, \ldots, n$, $j = 1, \ldots, 5$. The values of the covariates are generated from a standard normal distribution and regression parameters are fixed at: $\beta_1 = 0.5$, $\beta_2 = 0.4$, and $\beta_3 = 0.6$. The parameter ϕ is given 40. The random effects is generated by using Formula 8, when Z_0 and Z_1 are independent random variables that have standard normal distributions. Also we consider $\sigma_b^2 = 1$. Three data sets are generated by assuming $\lambda \in \{0, 2, 5\}$.

The Bayesian analysis is specified under priors (7). For each model, two parallel chains in different initial points were run, with 800,000 iterations for each chain. The results are presented considering the last 500,000 iterations. In addition, to avoid correlation problems in the generated chains, the thin value is considered 100.

In Table 1, random effects are assumed to follow normal and skew-normal distributions. Three data sets were generated by assuming $\lambda \in \{0, 2, 5\}$, representing behaviors from weak to strong skewness. To compare normal and skew-normal distributions for random effects, we calculate two criteria, the average of bias and the standard deviation (SD) that were computed as follow:

$$Biase(\hat{\theta}) = \bar{\hat{\theta}} - \theta_{True}, \qquad SD(\hat{\theta}) = \left(\frac{1}{n-1} \sum_{i=1}^{n} (\hat{\theta}_i - \bar{\hat{\theta}})^2 \right)^{1/2}$$

where $\hat{\theta}_i$ is the estimation of θ from the ith simulated data set.

In Table 1, the random effects are generated from skew-normal distribution for three cases, λ =0,2,5. Which λ =0 is the normal one. Then, the normal and the skew-normal distributions are allocated to random effects.

Table 1. Bias and standard deviations (SD) of parameter estimates under two distributions for random effects

		Normal		Skew-Normal	
Parameters	True Values	Bias	SD	Bias	SD
β_1	0.5	-0.008	0.011	0.151	0.152
β_2	0.4	0.001	0.001	0.005	0.031
β_3	0.6	-0.005	0.002	0.009	0.032
ϕ	40	0.125	3.531	0.213	3.213
p_0	0.1	-0.001	0.001	0.002	0.011
p_1	0.1	-0.001	0.001	0.002	0.014
σ_b^2	1	-0.0192	0.021	0.165	0.145
λ	0	-	-	0.21	0.198
β_1	0.5	0.154	0.250	0.037	0.155
β_2	0.4	0.008	0.035	0.005	0.021
β_3	0.6	0.011	0.026	0.003	0.023
ϕ	40	0.537	3.607	-0.237	3.602
p_0	0.1	-0.009	0.015	-0.003	0.012
p_1	0.1	-0.004	0.016	-0.001	0.014
σ_b^2	1	-1.489	1.525	0.004	0.481
λ	2	-	-	0.237	0.481
β_1	0.5	0.164	0.264	0.138	0.255
β_2	0.4	0.006	0.040	0.008	0.025
β_3	0.6	0.007	0.024	0.022	0.044
ϕ	40	-0.65	3.216	-0.266	1.921
p_0	0.1	-0.017	0.021	-0.031	0.033
p_1	0.1	-0.031	0.035	-0.022	0.025
σ_b^2	1	-6.327	6.399	-0.431	3.618
λ	5	-	-	0.799	1.147

Table 2. Bias and standard deviations (SD) of parameter estimates under different values of k on prior distribution λ, for λ=5

	k=2		k=3		k=4		k=5	
Parameters	Bias	SD	Bias	SD	Bias	SD	Bias	SD
β_1	0.138	0.255	0.141	0.258	0.143	0.301	0.148	0.311
β_2	0.008	0.025	0.007	0.027	0.008	0.029	0.008	0.031
β_3	0.022	0.044	0.023	0.043	0.024	0.042	0.026	0.047
ϕ	-0.266	1.921	-0.272	1.987	-0.285	2.142	-0.298	3.104
p_0	-0.031	0.033	-0.031	0.034	-0.033	0.033	-0.035	0.041
p_1	-0.022	0.025	-0.023	0.027	-0.026	0.029	-0.026	0.030
σ_b^2	-0.431	3.618	-0.491	4.011	-0.521	4.255	-0.558	4.861
λ	0.799	1.147	0.811	1.151	0.821	1.899	1.118	1.614

For λ =2,5 by comparing biase's and SD's, the skew-normal for random effects is better than of the normal distribution especially when λ=5. When λ =0 by comparing the biase's and the SD's of parameter estimates, the results show that although the normal one has a little better performance but the skew-normal distribution is also admitted for random effects. But when $\lambda >$ 0, allocation of a normal distribution to random effects is not desirable and it declines by increasing λ. Generally, our evaluation suggests that the skew-normal distribution for random effects is better than normal distribution. In other words, allocating skew-normal distribution to the random effects provide better estimates for the model parameters.

In addition, since the Bayesian analysis is based on prior distributions, $\lambda \sim HN(0, 10^k)$, we performed the Bayesian inference with different values of k. The results corresponding to the posterior analysis for λ=5 and $k = 2, \ldots, 5$ are reported in Table 2. Changing values of k for $k \geqslant 2$ shows that results are not sensitive to the values of k. Also, we allowed $\sigma_b^2 \sim I\Gamma(a, a)$ and $\phi \sim \Gamma(a, a)$, where $a \in \{0.01, 0.1\}$. The results show that the parameter estimates are not sensitive for values of a. In order to see how stable estimates are, two parallel chains with different initial values should be used. The convergence of the MCMC samples of parameters were verified by Gelman and Rubin test. Furthermore, for each parameter, we checked that the convergence is achieved for each chain.

6 Application

In this section, we apply our proposed model for analysis of LFS data that is a quarterly survey of households for measuring the economically active population conducted by the Statistical Centre of Iran. The LFS consists of probabilistic samples of household units that produces quarterly labor force and related estimates for all members of private settled households whose place of usual residence was located in Iran at the time of the enumeration. The sample rotation follows a 2-2-2 pattern. Therefore, each unit is considered as one household. A household unit is interviewed in two consecutive quarters and not in the sample for the next two quarters. It returns to the sample again for the next two quarters.

The LFS interview is divided into two parts: 1- household and demographic information, 2- labor force information. The second part for each 10 years or older person in the household is obtained. One of the main goals of labor force information is to classify persons as employed, unemployed or inactive. So the proportion of employed persons in the household can be calculated. We consider a panel data on 159 households who had been interviewed in the first and second quarters of 2013 and also the same quarters of 2014 in Tehran, the capital of Iran. Therefore, each household is interviewed four times, consisting 636 observations of 159 households. The definitions and concepts used in LFS were consistent with international recommendations of the International Labor Organisation in 1998.

Figure 1 displays the histogram of the proportion of employed persons in the household based on 159 households for each quarter separately that drop in the closed interval $[0, 1]$, where 0 and 1 respectively, represent households with no employed member and households with all members employed. Therefore, we have used the augmented mixed Beta regression model for which random effects have two different normal and skew-normal distributions. Based on the augmented mixed beta regression model, the prior distributions in (7) and logit link function (6), we considered the following structure to this panel data:

$$
\begin{aligned}
&Y_{ij}|\mu_{ij} \sim AB(p_0, p_1, \mu_{ij}\phi, (1-\mu_{ij})\phi),\\
&\text{logit}(\mu_{ij}) = \beta_1 + \beta_2 x_{1ij} + \beta_3 x_{2ij} + b_i, \quad i = 1, \ldots, 159,\ j = 1, 2, 3, 4,\\
&b_i \sim N(0, \sigma_b^2) \quad \text{or} \quad b_i \sim SN(0, \sigma_b^2, \lambda).
\end{aligned}
$$

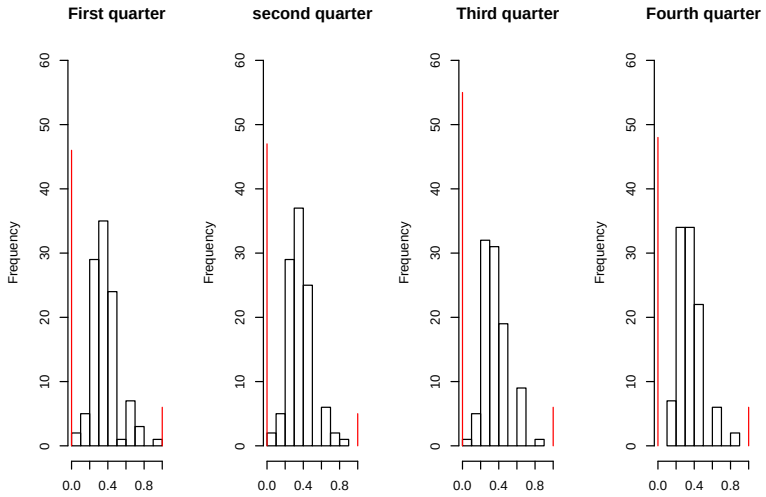

Figure 1. Histogram of the proportion of employed persons in the household based on the Labor Force Survey for each quarter separately for Tehran city data.

The density of Y_{ij}, $i = 1, \ldots, 159$, $j = 1, 2, 3, 4$ is given by:

$$f(y_{ij}; p_0, p_1, \mu_{ij}, \phi) = \begin{cases} p_0 & y_{ij} = 0 \\ p_1 & y_{ij} = 1 \\ (1 - p_0 - p_1) f(y_{ij}; \mu_{ij}, \phi) & y_{ij} \in (0, 1), \end{cases}$$

where $f(y_{ij}; \mu_{ij}, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu_{ij}\phi)\Gamma((1-\mu_{ij})\phi)} y_{ij}^{\mu_{ij}\phi - 1}(1-y_{ij})^{(1-\mu_{ij})\phi - 1}$. In this model, y_{ij} is the proportion of employed individuals in the household, p_0 is the probability of households with no employed member, p_1 is the probability of households with all members employed, x_{1ij} is the number of persons that have 10 years or older in the household and x_{2ij} is the household size. Also μ_{ij} is the mean of the response variable, ϕ is the precision parameter and b_i is a random effects which considers the potential correlation between responses over time. Hence significance of its variance shows the existence of correlated responses over time.

It should be noted that multicolinearity for these variables is investigated and the problem was not seen.

Considering prior densities in (7), we generated two parallel independent

Table 3. Model comparison for LFS data set

Model	LPML	EAIC	EBIC
Normal	-30.561	-28.76	-4.21
Skew-Normal	-25.32	-49.13	-24.58

Table 4. Parameter estimates (Est.) and standard deviations (SD) obtained after fitting normal or skew-normal distributions to random effects in LFS data sets. The 95% credible intervals are provided only for the desirable model.

	Skew-Normal				Normal	
Parameters	Est.	SD	2.5%	97.5%	Est.	SD
β_1	0.384	0.141	0.105	0.624	0.657	0.179
β_2	0.120	0.048	0.067	0.234	0.141	0.069
β_3	-0.500	0.063	-0.602	-0.416	-0.484	0.101
ϕ	43.092	3.323	36.642	45.287	38.734	8.575
p_0	0.309	0.018	0.273	0.322	0.309	0.018
p_1	.038	0.008	0.024	0.043	0.038	0.008
σ_b^2	0.007	0.004	0.003	0.009	0.156	0.113
λ	0.547	0.049	0.452	0.579	-	-

runs of Gibbs sampler chain with size 800,000 iterations for each parameter, discarding the first 500,000 iterations and avoiding correlation problems, we considered the thin value 100. To monitor the convergence of the Gibbs samples we used the methods presented by (Gelman and Rubin, 1992).

Table 3 shows the comparison among the two different models by using the selection criteria considered in Section 4. Notice that, the model with skew-normal distribution for random effects outperforms the one with normal distribution for all criteria.

Table 4 gives the parameter estimates (Est.) and standard deviations (SD) obtained from fitting models with random effects under skew-normal and normal distributions to the LFS data. The results show that the number of 10 year old or older persons in the household has a positive effect to explain the proportion of employed persons in the household. Reversely, the household size variable with a negative coefficient is included in the model. The estimate of precision parameter, $\hat{\phi} = 43.092$. The skewness parameter $\lambda = 0.547$, $p_0 = 0.309$ and $p_1 = 0.038$.

To monitor Markov chain convergence, we checked the PSRF of all pa-

Table 5. Parameter estimates (Est.) and standard deviations (SD) of parameter estimates under different values of "a" on prior distributions.

	a=0.01		a=0.1	
Parameters	Est.	SD	Est.	SD
β_1	0.384	0.141	0.411	0.151
β_2	0.120	0.048	0.115	0.051
β_3	-0.500	0.063	-0.487	0.071
ϕ	43.092	3.323	44.215	3.418
p_0	0.309	0.018	0.311	0.020
p_1	.038	0.008	.040	0.009
σ_b^2	0.007	0.004	0.009	0.004
λ	0.547	0.049	0.563	0.052

rameters. In MCMC method, we allocated two parallel chains with different starting values in 800,000 iterations. Then, we discarded the first 500,000 iterations. For each model, all PSRF are close to 1 for each parameters.

A sensitivity analysis was performed on the prior assumptions and initial values. We allowed $\sigma_b^2 \sim I\Gamma(a,a)$ and $\phi \sim \Gamma(a,a)$, where $a \in \{0.01, 0.1\}$. The results in Table 5 show that the parameter estimations are not very sensitive on the selected values of a. We also checked the sensitivity of the posterior estimates of β on different priors. The results in Figure 2 show that they are not sensitive on t-student priros with degrees of freedom (df) greater than 5.

7 Conclusion

The augmented Beta regression is utilizable to model data that are seen on $[0,1)$, $(0,1]$ or $[0,1]$. In this paper, we have considered the augmented Beta regression model with random effects under the skew-normal distributional assumption. Since the model was complicated, the Bayesian method was used through the Markov chain Monte Carlo algorithm. All the Bayesian computation were performed using accessible software WinBUGS. To show the performance of our proposed model, some intensive simulation studies have been done. In the application section, by Labor Force Survey data set, the proportion of employed persons in the household has been analyzed where the augmented mixed Beta with skew-normal random effects have been used.

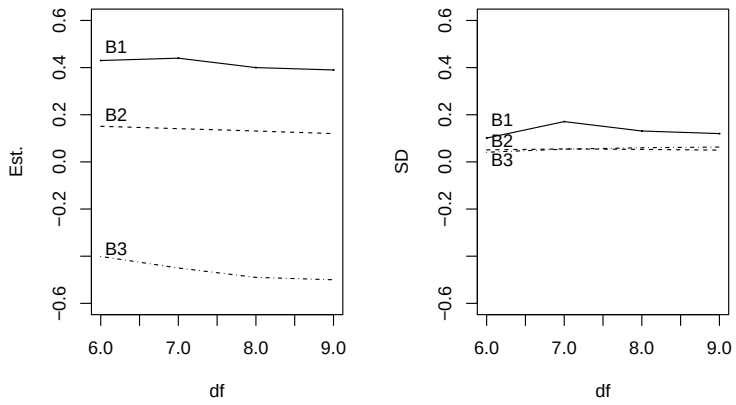

Figure 2. Parameter estimates (Est.) and standard deviations (SD) obtained for the parameters β_1, β_2 and β_3 with t-student priros with df=6,...,9

Since random effects on real data is generally not symmetric, the proposed model can be more appropriate model.

In this paper we used a random effects model under skew-normal distribution to analyzed the longitudinal proportion data, which is the most famous approach for considering association among repeated measures but one can consider marginal or transition model for analyzing such data as another future work.

Acknowledgement

The authors are thankful to the referees for their many helpful comments that greatly improved this paper. We also wish to acknowledge for the support from Center of Excellence in Spatial and Temporal Data Analysis of Tarbiat Modares University.

References

Arellano-Valle, R.B., Bolfarine, H. and Lachos, V.H. (2005). Skew-Normal Linear Mixed Models. *Journal of Data Science*, **3**, 415-438.

Azzalini, A. (1985). A Class of Distributions which Includes the Normal Ones. *Scandinavian Journal of Statistics*, **12**, 171–178.

Azzalini, A. and Dalla Valle, A. (1996). The Multivariate Skew-Normal Distribution. *Biometrika*, **83**, 715-726.

Carlin, B. and Louis, T. (2008). *Bayesian Methods for Data Analysis (Texts in Statistical Science)*. Chapman and Hall/CRC Press, New York.

Cribari-Neto, F. and Zeilis, A. (2010). Beta Regression In R. *Journal of Statistical Software*, **34**, 1-24.

Ferrari, S. and Cribari-Neto, F. (2004). Beta Regression for Modeling Rates and Pproportions. *Journal of Applied Statistics*, **31**, 799-815.

Figueroa-Zúñiga, Arellano–Valle, R.B. and Ferrari, S.L.P. (2013). Mixed Beta Regression: A Bayesian Perspective. *Computational Statistics and Data Analysis* , **61**, 137-147.

Galvis, M. D, Dipankar, B. and Victor, H. L. (2014). Augmented Mixed Beta Regression Models for Periodontal Proportion Data. *Statistics In Medicine*, **33**, 3759-3771.

Gelfand, A.E. and Dey, D. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society*, Series B. **56**, 501-514.

Gelman, A. and Rubin, D.B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, **7**, 457-472.

Hall, D.B. and Berenhaut, K.S. (2002). Score Tests for Heterogeneity and Overdispersion in Zero-Inflated Poisson and Binomial Regression Models. *J. Statist.*, (**7**), 415-430.

Lachenbruch, P.A. (2002). Analysis of Data with Excess Zeros. *Stat. Methods Med.*, Res. 11, (**4**), 297-302.

Lachos, V.H., Ghosh, P. and Arellano-Valle, R.B. (2010). Likelihood based Inference for Skew-Normal Independent Linear Mixed Models. *Statistica Sinica*, 303-322.

Lin, T.I. and Lee, J.C. (2008). Estimation and Prediction in Linear Mixed Models with Skew Normal Random Effects for Longitudinal Data. *Statistics in Medicine*, **9**, 1490-1507.

Ospina, R. and Ferrari, S. (2010). Inflated Beta Distributions. *Statist. Papers*, **23**, 193-213.

Verkuilen, J. and Smithson, M. (2012). Mixed and Mixture Regression Models for Continuous Bounded Responses Using the Beta Distribution. *Journal of Educational and Behavioral Statistics*, **37**, 82-113.

Zhang, D. and Davidian, M. (2001). Linear Mixed Models with Flexible Distributions of Random Effects for Longitudinal Data. *Biometrics*, **57**, 795-802.

Zohreh Fallah Mohsenkhani
Department of Statistics,
Tarbiat Modares University,
and Statistical Research And Training Center,
Tehran, Iran.
email: *zohrehf@yahoo.com*

Mohsen Mohammadzadeh
Department of Statistics,
Tarbiat Modares University,
Tehran, Iran.
email: *mohsen_m@modares.ac.ir*

Taban Baghfalaki
Department of Statistics,
Tarbiat Modares University,
Tehran, Iran.
email: *t.baghfalaki@modares.ac.ir*