Communications for Statistical Applications and Methods
2020, Vol. 27, No. 2, 189–199

https://doi.org/10.29220/CSAM.2020.27.2.189
Print ISSN 2287-7843 / Online ISSN 2383-4757

Bayesian inference for an ordered multiple linear regression with skew normal errors

Jeongmun Jeong[a], Younshik Chung[1,a]

[a]Department of Statistics, Pusan National University, Korea

Abstract

This paper studies a Bayesian ordered multiple linear regression model with skew normal error. It is reasonable that the kind of inherent information available in an applied regression requires some constraints on the coefficients to be estimated. In addition, the assumption of normality of the errors is sometimes not appropriate in the real data. Therefore, to explain such situations more flexibly, we use the skew-normal distribution given by Sahu *et al.* (*The Canadian Journal of Statistics*, **31**, 129–150, 2003) for error-terms including normal distribution. For Bayesian methodology, the Markov chain Monte Carlo method is employed to resolve complicated integration problems. Also, under the improper priors, the propriety of the associated posterior density is shown. Our Bayesian proposed model is applied to NZAPB's apple data. For model comparison between the skew normal error model and the normal error model, we use the Bayes factor and deviance information criterion given by Spiegelhalter *et al.* (*Journal of the Royal Statistical Society Series B (Statistical Methodology)*, **64**, 583–639, 2002). We also consider the problem of detecting an influential point concerning skewness using Bayes factors. Finally, concluding remarks are discussed.

Keywords: Bayes factor, deviance information criterion, influential point, Markov chain Monte Carlo, ordered linear regression, Skew normal distribution

1. Introduction

There are many applications with the statistical structures for which a constrained linear regression model is proper. It is often reasonable that the incomplete information available in an applied regression require some different types of constraints on their parameters to be estimated. For example, the amount of apples naturally increases as an apple tree grows older. However, if there is not enough data, the analysis results may differ from common sense. In order to prevent this, it is recommended to place an ordinal constraint. For these points, Chen and Deely (1996) considered one such application and illustrated why the Bayesian model is both attractive and appropriate. They then used the normal distribution for error terms in a constrained linear multiple regression model. But, the normality assumption of the errors is not appropriate in real data because some data sometimes have heavy tails or are asymmetric in various fields. To overcome this, as expected, the assumption of the error terms would be more flexible. Therefore, we use the skew normal distribution for error terms containing the normal distribution for the Bayesian ordered multiple linear regression model as a flexible distribution.

The skew normal distribution was first introduced by O'Hagan and Leonard (1976). Azzalini (1985) conducted study on the construction of the family of univariate skew-normal distributions.

[1] Corresponding author: Department of Statistics, Pusan National University, 2, Busandaehak-ro 63beon-gil, Geumjeong-gu, Busan 46241, Korea. Email: yschung@pusan.ac.kr

Published 31 March 2020 / journal homepage: http://csam.or.kr

Azzalini and Dalla-Valle (1996) extended the univariate skew-normal distribution to the multivariate case. The skew-normal distribution is a family of distributions with an additional parameter of bias. Some applications of the skew-normal regression are presented in Azzalini and Capitanio (1999), under a classical method. Sahu *et al.* (2003) proposed a new class of multivariate skew distributions to Bayesian regression models. Jang *et al.* (2009) also applied the skew elliptical distribution to Bayesian meta analysis. Jung *et al.* (2018) recently studied the Bayesian change-point problem with hierarchical prior distribution using skew distribution.

Our goal is to propose the Bayesian inference for an ordered multiple regression model with skew-normal error terms. We then show that Bayesian methodology is particularly suited to the constrained problem compared to frequentist methodology. Here, we use the Markov chain Monte Carlo (MCMC) method to resolve complicated integration problems related to Bayesian inference. We also perform simulation studies and apply our proposed model to New Zealand apple data that was already used by Chen and Deely (1996). We verify the convergence of MCMC for all parameters based on the Gelman-Rubin shrinkage factor. For model comparison between normal error model and skew normal error model, we use the Bayes factor (BF) and deviance information criterion (DIC) given by Spiegelhalter *et al.* (2002). We also consider the problem of detecting an influential point concerning skewness using Bayes factors.

The paper is organized as follows. In Chapter 2, we review the skew normal distribution and an ordered multiple linear regression. In Chapter 3, we discuss a Bayesian ordered multiple regression model with skew normal errors and its Bayesian computations using the MCMC method. The propriety of the associated posterior density is also shown under the given improper priors. We present the model comparison using BF and DIC. The measure K_d of the effect of observations on BF is employed for detecting influential data. In Chapter 4, we conduct simulation studies to check the performance of the skew-normal error model and apply our model to NZAPB's apple data. We then compare our results with Chen and Deely (1996)'s results. Concluding remarks are given in Chapter 5.

2. Ordered multiple linear regression with skew normal distribution

2.1. Skew normal distribution

According to Azzalini (1985), a random variable Z has a standard skew-normal distribution with probability density function given by

$$f_Z(z|\lambda) = 2\phi(z)\Phi(\delta z), \tag{2.1}$$

where z $\in (-\infty, \infty)$, ϕ, and Φ denote the probability density function and the distribution function of a standard normal distribution, respectively and δ is a real-valued parameter controlling the skewness of the distribution. In particular, one revisits the normal case if $\delta = 0$. If a location parameter μ and a scale parameter σ exist, Sahu *et al.* (2003) proposed the skew normal distribution SN(μ, σ^2, δ) as follows:

$$f(\epsilon|\mu, \sigma^2, \delta) = \frac{2}{\sqrt{\sigma^2 + \delta^2}}\phi\left[\frac{\epsilon - \mu}{\sqrt{\sigma^2 + \delta^2}}\right]\Phi\left[\frac{\delta}{\sigma}\frac{\epsilon - \mu}{\sqrt{\sigma^2 + \delta^2}}\right], \tag{2.2}$$

where $\epsilon \in (-\infty, \infty)$. Similarly, if $\delta = 0$, the skew normal distribution in (2.2) is also the same as the $N(\mu, \sigma^2)$.

2.2. Ordered multiple linear regression

First, we introduce the general expression of linear regression model as:

$$Y = E(Y|X) + \boldsymbol{\epsilon}$$

and

$$E(Y|X) = X\boldsymbol{\beta},$$

where $Y = (y_1, \ldots, y_n)'$ is response variable, $X = (x_1, \ldots, x_n)'$ is a $n \times p$ design matrix, $x_i = (x_{i1}, \ldots, x_{ip})'$ for $i = 1, \ldots, n$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, p is the number of parameters and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ denotes the error term which is usually distributed to the multivariate normal distribution. Denote the constrained set of parameter $\boldsymbol{\beta}$ by

$$\boldsymbol{S} = \left\{(\beta_1, \beta_2, \ldots, \beta_p)' : \boldsymbol{\beta} \in \boldsymbol{Q} \in \boldsymbol{R}^p\right\},$$

where $\boldsymbol{Q}$ is a subspace of S under some type of constraints and R^p is p-dimensional Euclidean space. Denote the constrained set of parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ by

$$\boldsymbol{\beta} \in Q = \left\{\boldsymbol{\beta} \in \boldsymbol{R^p}; 0 \le \beta_1 \le \beta_2 \le \cdots \le \beta_p\right\} \tag{2.3}$$

which is considered one of the constrained forms in the Bayesian main structure in Section 3 according to the real data in Section 4.

3. Bayesian Inference

Our model to be considered as:

$$Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad 0 \le \beta_1 \le \beta_2 \le \cdots \le \beta_p, \tag{3.1}$$

where $Y = (y_1, \ldots, y_n)'$ is response variable, $X = (x_1, \ldots, x_n)'$ is a $n \times p$ design matrix, $x_i = (x_{i1}, \ldots, x_{ip})'$, $i = 1, \ldots, n$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, p is the dimension of parameters and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ is distributed to $\mathrm{SN}(\mu, \sigma^2, \delta)$ in (2.2).

3.1. Bayesian model

According to Sahu *et al.*'s expression, it follows from (2.3) and (3.1) that the likelihood function of $\boldsymbol{\beta}, \sigma^2, \delta$ based on the data $\boldsymbol{y}, X$ is

$$\begin{aligned} L\left(\boldsymbol{\beta}, \sigma^2, \delta|\boldsymbol{y}, X\right) &= \prod_{i=1}^{n}\left[\frac{2}{\sqrt{\sigma^2+\delta^2}}\phi\left(\frac{y_i - \sum_{j=1}^{p} x_{ij}\beta_j}{\sqrt{\sigma^2+\delta^2}}\right)\Phi\left(\frac{\delta}{\sigma}\frac{y_i - \sum_{j=1}^{p} x_{ij}\beta_j}{\sqrt{\sigma^2+\delta^2}}\right)\right] \\ &= \prod_{i=1}^{n}\left[\frac{2}{\sqrt{\sigma^2+\delta^2}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2(\sigma^2+\delta^2)}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2\right)\right] \\ &\quad \times \prod_{i=1}^{n}\left[\int_{-\infty}^{\infty} I\left(w < \frac{\delta}{\sigma}\frac{y_i - \sum_{j=1}^{p} x_{ij}\beta_j}{\sqrt{\sigma^2+\delta^2}}\right)\frac{1}{\sqrt{2\pi}}e^{-\frac{w^2}{2}}dw\right]. \end{aligned} \tag{3.2}$$

We consider a noninformative priors on $\boldsymbol{\beta}$ based on Chen and Deely (1996). Then, the prior on $\boldsymbol{\beta}$ is taken as

$$\pi_1(\boldsymbol{\beta}) \propto \mathbf{1}_{\boldsymbol{Q}}(\boldsymbol{\beta}), \tag{3.3}$$

where $\mathbf{1}_{\boldsymbol{Q}}(\boldsymbol{\beta}) = 1$ if $\boldsymbol{\beta} \in \boldsymbol{Q}$, 0 otherwise. Finally, the priors of σ^2 and δ are given by

$$\pi_2\left(\sigma^2\right) \propto \frac{1}{(\sigma^2)^{k+1}} e^{-\frac{k}{\sigma^2}} \tag{3.4}$$

and

$$\pi_3(\delta) = N\left(\mu_\delta, \sigma_\delta^2\right), \tag{3.5}$$

respectively. Here, as k goes to zero, $\pi_2(\sigma^2)$ approached to $1/\sigma^2$ which is considered as the improper prior of σ^2. Since the prior distributions of each parameter are assumed to be independent, the joint prior density of $(\boldsymbol{\beta}, \sigma^2, \delta)$ is given by

$$\pi\left(\boldsymbol{\beta}, \sigma^2, \delta\right) \propto \pi_1\left(\boldsymbol{\beta}\right) \pi_2\left(\sigma^2\right) \pi_3(\delta). \tag{3.6}$$

Theorem 1. *Suppose that the priors $\boldsymbol{\beta}, \sigma^2$, and δ are given in (3.3)–(3.6). Then, the posterior density of $\boldsymbol{\beta}, \sigma^2$, and δ based on the likelihood in (3.2) is proper if $n > p$.*

We provide a proof of Theorem 1 in the Appendix.

3.2. Bayesian computations

Here, it is hard to handle the integration part to calculate full conditional density of each parameter for Bayesian computation. Using the data augmentation suggested by Tanner and Wong (1987), we consider $\mathbf{z} = (z_1, z_2, \ldots, z_n)$ as unobserved data in (3.2). Then the complete likelihood function of $\boldsymbol{\beta}, \sigma^2, \delta$, and $\boldsymbol{z}$ based on the data $\boldsymbol{y}$ and X is given by

$$\begin{aligned} L_c\left(\boldsymbol{\beta}, \sigma^2, \delta, \boldsymbol{z} | \boldsymbol{y}, X\right) = \prod_{i=1}^{n} \left[I\left(z_i < \frac{\delta}{\sigma} \frac{y_i - \sum_{j=1}^{p} x_{ij}\beta_j}{\sqrt{\sigma^2 + \delta^2}} \right) \right] \left(\frac{1}{\sqrt{\sigma^2 + \delta^2}} \right)^n \\ \times \exp\left(-\frac{1}{2} \frac{1}{\sqrt{\sigma^2 + \delta^2}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \sum_{i=1}^{n} z_i^2 \right), \end{aligned} \tag{3.7}$$

where z $\in (-\infty, \infty)$. From (3.6) and (3.7), the complete posterior density function of $(\beta, \sigma^2, \delta, \boldsymbol{z})$ is given by

$$\pi\left(\boldsymbol{\beta}, \delta, \sigma^2, \boldsymbol{z} | \boldsymbol{y}, X\right) \propto L_c\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{z} | \boldsymbol{y}, X\right) \pi\left(\boldsymbol{\beta}, \delta, \sigma^2\right)$$

subject to $0 \le \beta_1 \le \beta_2 \le \cdots \le \beta_p < \beta_{p+1} = \infty$.

To sample from the complete posterior distribution $\pi(\boldsymbol{\beta}, \delta, \sigma^2, \boldsymbol{z} | \boldsymbol{y}, X)$ given in (3.7), we can apply the Gibbs sampler by taking the full conditional densities as:
For $i = 1, \ldots, n$, let $z_{(-i)} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$. Then, the full conditional density of z_i is given by

$$p\left(z_i | \boldsymbol{\beta}, \sigma^2, \delta, \boldsymbol{z_{(-i)}}, \boldsymbol{y}, X\right) \propto I\left(z_i < \frac{\delta}{\sigma} \frac{y_i - \sum x_{ij}\beta_j}{\sqrt{\sigma^2 + \delta^2}} \right) \exp\left(-\frac{1}{2} z_i^2 \right) \tag{3.8}$$

which is the truncated normal distribution $TN(0,1)$ on $-\infty \le z_i \le (\delta/\sigma)\{(y_i - \sum x_{ij}\beta_j)/\sqrt{\sigma^2+\delta^2}\}$.

Define $\beta_{(-j)} = (\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p)$. Then, for $j = 1, \ldots, p$, the full conditional density of β_j is given by

$$p\left(\beta_j|\beta_{(-j)}, \sigma^2, \delta, \boldsymbol{z}, \boldsymbol{y}, X\right) = TN\left(\mu_{\beta_j}, \sigma^2_{\beta_j}\right) \tag{3.9}$$

subject to $\beta_{j-1} \le \beta_j \le \beta_{j+1}$ with $\beta_0 = 0$, where $\mu_{\beta_j} = (\sum_{i=1}^n x_{ij}^2)^{-1} \sum_{i=1}^n (y_i - \sum_{l\neq j} x_{il}\beta_l)x_{ij}$ and $\sigma^2_{\beta_j} = (\sigma^2+\delta^2)/\sum_{i=1}^n x_{ij}^2$.

Next, the full conditional density of σ^2 is given by

$$p\left(\sigma^2|\boldsymbol{\beta}, \delta, \boldsymbol{z}, \boldsymbol{y}, X\right) \propto \left(\sigma^2+\delta^2\right)^{-\frac{n}{2}} \left(\sigma^2\right)^{-(k+1)} \exp\left(-\frac{k}{\sigma^2}\right) \exp\left(-\frac{1}{2(\sigma^2+\delta^2)} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij}\beta_j\right)^2\right). \tag{3.10}$$

Finally, the full conditional density of δ is given by

$$p\left(\delta|\boldsymbol{\beta}, \sigma^2, \boldsymbol{z}, \boldsymbol{y}, X\right) \propto \left(\sigma^2+\delta^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2(\sigma^2+\delta^2)} \sum_{i=1}^n \left(y_i - \sum_{j=1}^p x_{ij}\beta_j\right)^2 - \frac{1}{2\sigma^2_\delta}(\delta-\mu_\delta)^2\right). \tag{3.11}$$

It can be seen that the full conditional distributions in (3.8) and (3.9) are the truncated normal distibution. Therefore, z_i's and β_i's can be easily generated by Robert's (1995) method. However, both full conditional distributions of σ^2 and δ in (3.10) and (3.11) cannot be reduced analytically to well-known distributions. Therefore, it is impossible to sample directly by the standard methods. As suggested by Chib and Greenberg (1995), a hybrid MCMC algorithm is used by combined a Metropolis-Hastings sampling and Gibbs sampler using the suitable proposed distributions. The proposal distribution of σ^2 is inverse gamma distribution with shape parameter $n/2$ and scale parameter $\sum_{i=1}^n (y_i - x_i'\boldsymbol{\beta})^2$ which is the same as full conditional density of σ^2 under normal error model, where $x_i = (x_{i1}, \ldots, x_{ip})'$. Finally because the support of δ is the full set of real number, we set the proposal distribution of δ as normal distribution with mean $\delta^{(k-1)}$ and variance σ^2_δ. The hybrid MCMC algorithm is working as:

Step 1: Start with initial point $(\boldsymbol{z}^{(0)}, \boldsymbol{\beta}^{(0)}, \sigma_2^{(0)}, \delta^{(0)})$.

Step 2: Set $k = 1$.

Step 3: for $i = 1, \ldots, n$, generate $z_i^{(k)}$ from $TN(0,1)$ on $-\infty \le z_i \le (\delta/\sigma)\{(y_i - \sum x_{ij}\beta_j)/\sqrt{\sigma^2+\delta^2}\}$ in (3.8).

Step 4: for $j = 1, \ldots, p$, generate $\beta_j^{(k)}$ from $TN(\mu_{\beta_j}, \sigma^2_{\beta_j})$ subject to $\beta_{j-1} \le \beta_j \le \beta_{j+1}$ in (3.9) where $\mu_{\beta_j} = (\sum_{i=1}^n x_{ij}^2)^{-1} \sum_{i=1}^n (y_i - \sum_{l\neq j} x_{il}\beta_l)x_{ij}$ and $\sigma^2_{\beta_j} = (\sigma^2+\delta^2)/\sum_{i=1}^n x_{ij}^2$.

Step 5: Using Metropolis-Hastings algorithm, generate $\sigma^{2(k)}$ from $IG(n/2, \sum_{i=1}^n (y_i - x_i'\boldsymbol{\beta})^2)$ as proposal distribution in (3.10).

Step 6: Using Metropolis-Hastings algorithm, generate $\delta^{(k)}$ from normal distribution $N(\delta^{(k-1)}, \sigma^2_\delta)$ as proposal distribution in (3.11).

Step 7: $k = k + 1$.

Step 8: Repeat Steps 3–7, T times.

Here, $IG(a, b)$ denotes the inverse gamma distribution with shape parameter a and scale parameter b. Then, for example, the posterior mean of β_j in (3.9) is approximated by

$$\hat{\beta}_j = \frac{1}{K(T-S)} \sum_{k=1}^{K} \sum_{i=S+1}^{T} \beta_{jk}^{(i)},$$

where S is the burn-in period and K is the number of parallel chains. Here, we decide the convergence to have been reached after S iterations of and MCMC algorithm have been performed on K multiple chains depending on Gelman-Rubin (1992)'s shrinkage factor. Then for each chain k, the observations $\beta_{jk}^{(1)}, \beta_{jk}^{(2)}, \ldots, \beta_{jk}^{(S)}$ are discarded and $\beta_{jk}^{(i)}$, $S+1 \leq i \leq T$ are considered as an independent sample from the posterior distribution of the Markov chain, which is typically the posterior distribution.

3.3. Model comparisons

Generally, for testing M_0: skew normal error model versus, M_1: normal error model, Bayes factor (BF_{01}) in favor of the model M_0 (or against the other model M_1) is used as:

$$\mathrm{BF}_{01} = \frac{f(X|M_0)}{f(X|M_1)}, \tag{3.12}$$

where $f(X|M_i) = \int f(X|\boldsymbol{\beta}, \sigma^2, \delta) f(\boldsymbol{\beta}, \sigma^2, \delta|M_i) d\boldsymbol{\beta} d\sigma^2 d\delta$ is the marginal likelihood of the model M_i. Then, if $\mathrm{BF}_{01} > 1$, then accept the model M_0 based on the given data. Newton and Raftery (1994) suggested that the marginal likelihood may be estimated by the harmonic mean of the likelihoods of a sample from the posterior distribution using MCMC. For notational conveniences, let $\theta = (\boldsymbol{\beta}, \sigma^2, \delta)$. Then, the Bayes factor can be estimated as:

$$\widehat{\mathrm{BF}}_{01} = \frac{\hat{f}(X|M_0)}{\hat{f}(X|M_1)}, \tag{3.13}$$

where $\hat{f}(X|M_j) = [(1/m) \sum_{i=1}^{m} f(X|\theta_j^{(i)})^{(-1)}]^{-1}$ and $\theta_j^{(i)}$ is the i^{th} parameter samples from MCMC in model M_j, $j = 0, 1$.

It is natural to use criterion based on trade-off between the fit of the data to the model and the corresponding complexity of the model. Therefore, we compare the models based on DIC given by Spiegelhalter *et al.* (2002) for Bayesian model comparison:

$$\mathrm{DIC} = \bar{D}(\theta) + p_D, \tag{3.14}$$

where $\bar{D}(\theta) = E[D(\theta)|\mathbf{y}, X]$, $p_D = \bar{D}(\theta) - D(\bar{\theta})$, $D(\theta) = -2 \log L(\theta|\mathbf{y}, X)$, and $\bar{\theta}$ is a posterior mean. Here, p_D is measure of complexity and the effective number of parameters. So, DIC is 'goodness-of-fit' plus 'the effective number of parameters'. Therefore model with smaller DIC is better than the other model with larger DIC based on the data.

3.4. Influential observation

To see the effect of single observation d on the Bayes factor, Pettit and Young (1990) suggested K_d as:

$$K_d = \log \mathrm{BF}_{01} - \log \mathrm{BF}_{01}^{(d)},$$

where BF_{01} is defined in (3.12), $\mathrm{BF}_{01}^{(d)} = f(X^{(d)}|M_0)/f(X^{(d)}|M_1)$ is the Bayes factor without a single observation d and $X^{(d)}$ is a full data except a d^{th} single observation. The larger values of $|K_d|$ indicates

that observation d has a large influence on the Bayes factor. This K_d is expressed as the difference in the logarithms of the conditional predictive ordinates (CPO) for the two models. That is,

$$\begin{aligned} K_d &= \log \frac{f(X|M_0)}{f(X|M_1)} - \log \frac{f(X^{(d)}|M_0)}{f\,(X^{(d)}|M_1)} \\ &= \log \frac{f(X|M_0)}{f\,(X^{(d)}|M_0)} - \log \frac{f(X|M_1).}{f\,(X^{(d)}|M_1)}. \end{aligned} \tag{3.15}$$

Pettit and Young (1990) suggested that an observation with $|K_d| > 1/2$ might be influential based on Jeffrey's scale of evidence for assessing Bayes factors. If K_d in (3.15) is positive, it means that

$$\log \frac{f(X|M_0)}{f(X^{(d)}|M_0)} > \log \frac{f(X|M_1)}{f(X^{(d)}|M_1)}.$$

Therefore, the difference between including and not including d^{th} observation in the M_0 is larger than that of M_1. So, the observation with $K_d > 1/2$ supports M_0 and the observation with $K_d < -1/2$ supports M_1.

4. Real data analysis

The New Zealand Apple and Pear Marketing Board (NZAPB) is a statutory authority that trades and manages every contract in exporting New Zealand apples. Therefore, the more than 1,500 apple growers in New Zealand engage in the global community as one grower making international trade contracts. It is significant to realize that the data recorded consist of total harvest for each grower and for each apple variety. However, the NZAPB has also recorded the number of trees at each age, for each grower and for each variety. For the purpose of NZAPB, ages of trees vary from 2 to 11, where a tree of age "11" means 11 or older and is considered to be a full-grown tree. The year 1 is not included in the analysis because its production is close to zero. This is a set of 207 records, each record consisting of the total amount of fruit produced and the number of trees of each age. See Chen and Deely (1996) for more details of NZAPB's apple data. It is reasonable that a linear regression model be used, where the quantity of fruit produced is regressed on the tree numbers at each age, beginning at physical age 2, for 10 years up to a full-grown tree. For notational convenience, we consider ages varying from 1 to 10, matching the association that age 1 is the physical age 2 and so on. Thus, we let, for $i = 1, \ldots, n = 207$,

$$Y_i = \sum_{j=1}^{10} \beta_j x_{ij} + \epsilon_i,$$

where $\epsilon_i \sim SN(0, \sigma^2, \delta)$ in (2.2), x_{ij} is number of trees at age j for grower i, β_j equals average number of cartons produced by trees at age j for $j = 1, \ldots, 9$, and β_{10} is average number of cartons produced by all full-grown trees. Furthermore, we note that β_i represents the average over all trees of age i. Chen and Deely (1996) mentioned that since growers do not usually allow poor trees to persist, they proposed the constrained multiple linear regression with normal error as:

$$0 \le \beta_1 \le \beta_2 \le \cdots \le \beta_{10},$$

which are called monotone ordered constraints.

Table 1: DICs for values of μ_δ and σ^2_δ

μ_δ	σ^2_δ		
	1	10	100
−1.0	2832.941	2832.997	2833.964
−0.5	2832.924	2832.906	2833.867
0.0	2832.792	2832.865	2833.693
0.5	2832.713	2832.799	2833.547
1.0	2832.670	2832.683	2833.444

DIC = deviance information criterion.

Table 2: Model comparison for apple data

Model	$\bar{D}(\theta)$	p_D	DIC
Normal error	2831.803	11.1669	2822.9699
Skew normal error	2731.566	51.1039	2782.6699

DIC = deviance information criterion.

Table 3: Normal error and skew normal error estimation

Parameter	Normal error		Skew normal error	
	Estimate	Standard error	Estimate	Standard error
β_1	0.01374	0.00806	0.01723	0.00019
β_2	0.02496	0.00820	0.01723	0.00019
β_3	0.17749	0.01113	0.18122	0.00016
β_4	0.31129	0.04111	0.29833	0.00158
β_5	0.55423	0.07618	0.55195	0.00601
β_6	0.78064	0.03463	0.81104	0.00126
β_7	0.81682	0.04071	0.81131	0.00129
β_8	0.93711	0.09977	0.86252	0.04474
β_9	1.04100	0.11945	0.89909	0.05384
β_{10}	1.22821	0.17487	1.29236	0.28248
σ^2	53469.3	5406.4	51267.8	5002.1
δ	·	·	2.03320	0.08099

We investigated the robustness of the ordered multiple linear regression with skew normal model according to the different values of μ_δ and σ^2_δ in (3.5) based on the values of DICs in (3.14). Table 1 shows that our proposed model is robust based on the DIC values.

To compare the skew normal model to the original normal model informally, compute BF_{01} in (3.13), the effective number of parameters p_D and the DIC in (3.14). Since $\widehat{\mathrm{BF}}_{01} = 10.73$, we accept the skew normal error model. Table 2 shows that the skewed model improved the original normal model based on the values of DICs. In particular, for the normal model, the effective number of parameters p_D is almost likely to the real number of parameters in the regression model.

Table 3 shows the Bayesian estimates of parameters of regressions, variation and skewness from rival models. All ten regression parameters β_i are significant in all models since the corresponding highest posterior density (HPD) intervals do not include the value zero. The Bayesian estimate σ^2 for the skew model is smaller than the normal model. This is expected since the variability and skewness are interchangeable to a certain extent. This means that the normal error model fails to control the skewness having larger variability. Table 2 shows that the result of Table 3 comes from the significance of the skew parameter. The skewness parameter δ is estimated to be positive in the skew normal model: it denotes that the given data fit to the right skewed normal mode mentioned previously. Moreover, δ is significant under the skew-normal model since the 95% HPD interval is

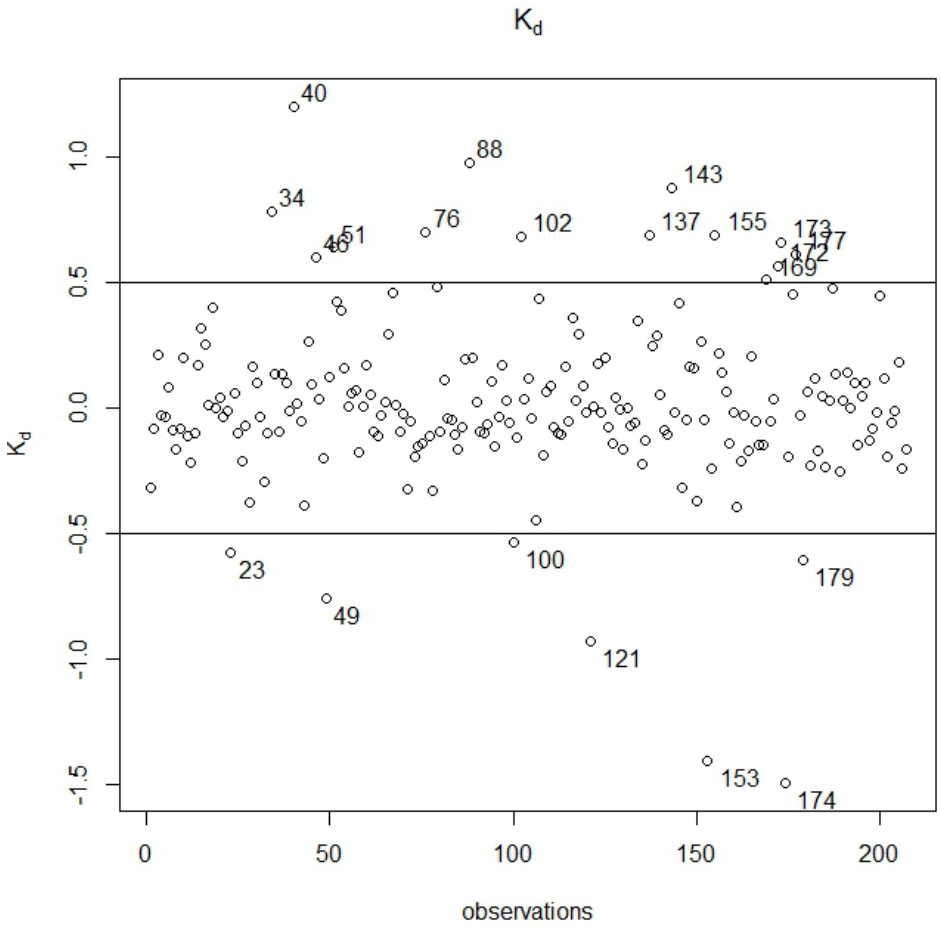

Figure 1: *K_d plot.*

(1.8743, 2.354). Thus, we can conclude that significant skewness is required to model the data.

In Figure 1, the plot K_d for normal error model and skew normal error model are displayed. There are 21 observations with $|K_d| > 1/2$. 14 observations support normal error model and 7 observations do skew normal error model. So, there are lots of the observations supporting normal error model.

5. Conclusions

In this paper, we presented a Bayesian ordered multiple linear regression with skew normal errors. For the case for hard to compute full conditional density, we use data augmentation method and expand the observed likelihood. Thus, we easily get estimates of parameters by MCMC. We show that the associated posterior density based on our Bayesian skew normal model is proper under the improper priors. That is, the marginal posterior density is finite.

We can detect which observation support normal error or skew normal error using the K_d measure. In the next study, we may consider the extension of the K_d measure in excluding two or more influence observations to check which group of observations are supported in some model.

Finally, we apply our model to data that has various types of constraint and an other flexible error-terms model.

Appendix: Proof of Theorem 1

We first consider the integral of the likelihood in (3.2) times the prior in (3.5). Let

$$A = \int \cdots \int L\left(\boldsymbol{\beta}, \sigma^2, \delta | \mathbf{y}, \mathbf{X}\right) \pi_1(\boldsymbol{\beta}) \pi_2\left(\sigma^2\right) \pi_3(\delta) d\boldsymbol{\beta} d\sigma^2 d\delta. \tag{A.1}$$

It is sufficient to show that A is finite. In the following derivation, the value of the constant C may not be the same from line to line :

$$\begin{aligned}
A &= C \int \cdots \int \left[\prod_{i=1}^{n} \frac{2}{\sqrt{\sigma^2 + \delta^2}} \phi\left(\frac{y_i - \sum_{j=1}^{p} x_{ij}\beta_j}{\sqrt{\sigma^2 + \delta^2}} \right) \Phi\left(\frac{\delta}{\sigma} \frac{y_i - \sum_{j=1}^{p} x_{ij}\beta_j}{\sqrt{\sigma^2 + \delta^2}} \right) \right] \times \pi_1(\boldsymbol{\beta}) \pi_2\left(\sigma^2\right) \pi_3(\delta) d\boldsymbol{\beta} d\sigma^2 d\delta \\
&\leq C \int \int \left(\sigma^2 + \delta^2\right)^{-\frac{n}{2}} \left[\prod_{i=1}^{p} \int_{\beta_{i-1}}^{\beta_{i+1}} \exp\left(-\frac{1}{2\left(\sigma^2 + \delta^2\right)} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right) d\beta_i \right] \times \pi_2\left(\sigma^2\right) \pi_3(\delta) d\sigma^2 d\delta \\
&\leq C \int \int \left(\sigma^2 + \delta^2\right)^{-\frac{n-p}{2}} \pi_2(\sigma^2) \pi_3(\delta) d\sigma^2 d\delta \\
&\leq C \int \int \left(\sigma^2\right)^{-\frac{n-p}{2}} \pi_2\left(\sigma^2\right) \pi_3(\delta) d\sigma^2 d\delta \\
&\leq C \int \left[\int \left(\sigma^2\right)^{-\frac{n-p}{2}+k+1} \exp\left(-\frac{k}{\sigma^2} \right) d\sigma^2 \right] \pi_3(\delta) d\delta
\end{aligned} \tag{A.2}$$

Then, the inner integral in the bracket of the equation (A.2) with respect σ^2 is finite if $n > p$. Therefore A is obviously finite since the prior $\pi_3(\delta)$ of δ is assumed to be normal density.

References

Azzalini A (1985). A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics*, **12**, 171–178.

Azzalini A and Capitanio A (1999). Statistical applications of the multivariate skew-normal distributions, *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, **61**, 579–602.

Azzalini A and Dalla-Valle A (1996). The multivariate skew-normal distribution, *Biometrika*, **83**, 715–726.

Chen MH and Deely JJ (1996). Bayesian analysis for a constrained linear multiple regression problem for predicting the new crop of apples, *Journal of Agricultural, Biological, and Environmental Statistics*, **1**, 467–489.

Chib B and Greenberg E (1995). Understanding the metropolis-hastings algorithm, *The American Statistician*, **49**, 327–335.

Gelman A and Rubin DB (1992). Inference from iterative simulation using multiple sequences, *Statistical Science*, **7**, 457–472.

Jang J, Chung Y, Kim C, and Song S (2009). Bayesian meta-analysis using skewed elliptical distributions, *Journal of Statistical Computation and Simulation*, **79**, 691-704.

Jung M, Song S, and Chung Y (2018). Bayesian change-point problem using Bayes factor with hierarchical prior distribution, *Communications in Statistics - Theory and Methods*, **46**, 1352–1366.

Newton MA and Raftery AE (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion), *Journal of the Royal Statistical Society. Series B (Methodology)*, **56**, 3–48.

O'Hagan A and Leonard T (1976). Bayes estimation subject to uncertainty about parameter constraints, *Biometrika*, **63**, 201–203.

Pettit LI and Young KDS (1990). Measuring the effect of observations on Bayes factors, *Biometrika*, **77**, 455–466.

Robert CP (1995). Simulation of truncated normal variables, *Statistics and Computing*, **5**, 121–125.

Sahu SK, Dey DK, and Branco MD (2003). A new class of multivariate skew distributions with applications to Bayesian regression models, *The Canadian Journal of Statistics*, **31**, 129–150.

Spiegelhalter DJ, Best NG, Carlin BP, and Van der Linde A (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, **64**, 583–639.

Tanner MA and Wong WH (1987). The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association*, **82**, 528–540.

Received July 13, 2019; Revised November 26, 2019; Accepted December 18, 2019