Skew-normal median regression model with applications

Xingyun Cao¹, Gege Wang¹, Liucang Wu*¹

¹Faculty of Science, Kunming University of Science and Technology, Kunming, China.

*Corresponding Author: Liucang Wu, email: wuliucang@163.com

Abstract. Skewed/asymmetrical data often appear in many research fields such as finance, economy, climate science, environmental science, engineering technology, social sciences and biomedicine. Little research has been carried out to establish median regression model to investigate the influence of some covariates on distributional characteristics. Since the median can capture the "medium" level of the skewed data, the main objective of this paper is to construct the median regression model for the analysis of skew-normal data. We develop an expectation-maximization (EM) algorithm facilitated by Newton-Raphson algorithm to estimate the regression coefficients and parameters. The proposed methods are evaluated by some Monte Carlo simulations and are illustrated by a real data analysis.

1. Introduction

In the real world, most of the collected data are skewed (or asymmetrical). The skewed distribution can reveal the asymmetric information when the observations are skewed as an alternative data analysis tool. Asymmetrical/skewed distributions have an additional shape parameter to represent the direction of the asymmetry of the density. If the skewness in observations is ignored, statistical inferences with symmetric distributions may result in biased or even misleading conclusions. In recent years, many skewed distributions including the skew-normal distribution [1] and the skew-t-normal distribution [2] were proposed for modeling such skewed observations.

Median is the value of "medium" level of the population. Its characteristic is that it is not affected by outliers. Especially for skewed distribution, the representation of the median is better than the mean. It has important practical significance in economics, social sciences and other disciplines. Many countries in the world use the median to describe the per capita income of residents, and the median is more robust and almost not affected by the income changes at the high and low ends.

The multivariate skew-normal distribution was developed by [3]-[5]. Especially, [6] investigated a classical and Bayesian approach for the skew-normal nonlinear regression model. For the measurement error data, [7] studied the parameter estimation of the skew-normal linear measurement error model. [8] considered the parameter estimation of the skew-normal linear mixed measurement error model. For mixture skewed data, [9] proposed a skew-normal mixture regression models. For variable selection, [10] investigated the simultaneously variable selection in joint location and scale models of the skew-normal distribution. [11] proposed variable selection in joint location, scale and skewness models based on the skew-normal distribution.

However, the above statistical inference methods are only limited to location models. Little research has been carried out to establish median regression model to investigate the influence of some covariates on distributional characteristics. The main objective of this paper is to construct the median regression model for the analysis of skew-normal data to capture the "medium" level. The rest of the paper is organized as follows. In Section 2, we introduce the skew-normal median regression model.
In Section 3, EM algorithm facilitated by Newton-Raphson algorithm to estimate the regression coefficients and parameters. In Section 4, we conduct some Monte Carlo simulations to evaluate the finite sample performance of the proposed method. In Section 5, the body mass index data of Australia are analyzed to illustrate the proposed methods. A discussion is given in Section 6.

2. Skew-Normal Median Regression Model

A random variable $Y$ is said to have a skew-normal distribution with location parameter $\mu$, scale parameter $\sigma$, and skewness parameter $\lambda$, denoted by $Y \sim SN(\mu, \sigma, \lambda)$, if its probability density function (pdf) is [1]

$$f_{SN}(y|\mu, \sigma, \lambda) = \frac{2}{\sigma} \phi \left( \frac{y-\mu}{\sigma} \right) \Phi \left( \lambda \frac{y-\mu}{\sigma} \right), \quad x \in \mathbb{R}$$  \hspace{1cm} (1)

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the pdf and cumulative distribution function (cdf) of the standard normal distribution, respectively. When $\lambda = 0$, the skew-normal distribution reduces to $N(\mu, \sigma^2)$. The distribution is right skewed if $\lambda > 0$ and is left skewed if $\lambda < 0$.

The expectation and variance of the random variable $Y$ are given by

$$\text{Mean}(Y) = E(Y) = \mu + \mu(\lambda), \quad \text{Var}(Y) = \sigma^2(\lambda)\sigma^2$$  \hspace{1cm} (2)

respectively, where

$$\mu(\lambda) = \frac{2}{\pi \sqrt{1 + \lambda^2}}, \quad \sigma^2(\lambda) = 1 - \frac{2\lambda}{\pi(1 + \lambda^2)}$$

From [12], we know that the mode of (1) can be expressed as

$$\text{Mode}(Y) = \mu + m_0(\lambda)\sigma$$  \hspace{1cm} (3)

where

$$m_0(\lambda) = \frac{\mu(\lambda) - \frac{t_0(\lambda)\sigma(\lambda)}{2} - \text{sgn}(\lambda)\frac{\sqrt{\pi}}{2} \exp \left( - \frac{2\pi}{|\lambda|} \right)}{\frac{2}{\lambda^2}}$$

We know that the mean, mode and median have the following quantitative relationship:

$$|\text{Mean} - \text{Mode}| \geq 3|\text{Mean} - \text{Median}|$$

Thus, we obtain

$$\text{Median}(Y) \approx \mu + \frac{\sigma}{3} \left[ m_0(\lambda) + 2\mu(\lambda) \right]$$  \hspace{1cm} (4)

The aim of this paper is to build regression model for $\text{Median}(Y)$ to capture the medium level of the skew-normal distribution. We consider the following median regression model:

$$Y_1 \sim SN(\mu_i, \sigma^2, \lambda), \quad \text{Median}(Y_1) = x_i\beta, \quad i = 1, \ldots, n.$$  \hspace{1cm} (5)

where $Y_1$ is the response variable, $x_i = (x_{i1}, \ldots, x_{ip})$ is the vector of covariates, $\beta = (\beta_1, \ldots, \beta_p)^T$ is an unknown vector of the median regression coefficients. From (4), we have

$$\mu_i \approx x_i\beta - \frac{\sigma}{3} \left[ m_0(\lambda) + 2\mu(\lambda) \right]$$  \hspace{1cm} (6)

3. MLEs of Parameters Via an EM Algorithm

Let $y_1$ be the realization of $Y_1 \sim SN(\mu_i, \sigma^2, \lambda)$, and $Y_{obs} = \{y_1\}_{i=1}^n$ denote the observed data and Let $\theta = (\beta^T, \sigma^2, \lambda)^T$ denote the vector of parameters in the median regression model. The observed-data log-likelihood function of $\theta$ for the model is given by

$$L(\theta|Y_{obs}) = c - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu_i)^2 + \sum_{i=1}^n \log[\Phi(\tau_i)]$$  \hspace{1cm} (7)

where $c$ is a constant, $\tau_i = \lambda(y_i - \mu_i)/\sigma$. The EM algorithm [13] can be used to calculate the MLEs of $\theta$, since the standard skew-normal distribution is a mixture of the standard normal distribution and the standard truncated normal distribution.

3.1. A mixture representation of the skew-normal distribution

The random variable $Y \sim SN(\mu, \sigma^2, \lambda)$ has the following stochastic representation (SR):
\[ Y = \mu + \sigma \left( \frac{\lambda |U|}{\sqrt{1 + \lambda^2}} + \frac{V}{\sqrt{1 + \lambda^2}} \right) \]  \hspace{1cm} (8)

where \( U, V \sim N(0,1) \). Let \( R = |U| \), then the SR (8) can be equivalently expressed in the following mixture representation:

\[ Y | (R = r) \sim N \left( \mu + \sigma \frac{\lambda r}{\sqrt{1 + \lambda^2}}, \frac{\sigma^2}{1 + \lambda^2} \right) \quad \text{and} \quad R \sim \text{TN}(0,1; (0, \infty)) \]  \hspace{1cm} (9)

where \( \text{TN}(\mu, \sigma^2; (a, b)) \) denotes the truncated normal distribution. So that the joint pdf of \((Y, R)\) is

\[ f_{(Y,R)}(y, r) = \frac{\sqrt{1 + \lambda^2}}{\pi \sigma} \exp \left\{ - \frac{1 + \lambda^2}{2\sigma^2} \left[ (y - \mu)^2 - \frac{2\sigma \lambda (y - \mu)r}{\sqrt{1 + \lambda^2}} + \sigma^2 r^2 \right] \right\} \]  \hspace{1cm} (10)

It is easy to show that

\[ R | (Y = y) \sim \text{TN} \left( \frac{\tau}{\sqrt{1 + \lambda^2}}, \frac{1}{1 + \lambda^2}; (0, \infty) \right), \quad \text{with} \quad \tau = \frac{\lambda (y - \mu)}{\sigma} \]

so that

\[ E(R | Y = y) = \frac{\tau}{\sqrt{1 + \lambda^2}} + \frac{1}{1 + \lambda^2} \Phi(\tau) \]  \hspace{1cm} (11)

\[ E(R^2 | Y = y) = \frac{\tau}{\sqrt{1 + \lambda^2}} + \frac{\tau^2}{1 + \lambda^2} E(R | Y = y) \]  \hspace{1cm} (12)

which will be used in the EM algorithm.

### 3.2. The EM algorithm

Based on the mixture representation (8), for each \( Y_i = y_i \), we introduce a latent variable \( R_i = r_i \) to form the complete data \( Y_{\text{com}} = Y_{\text{obs}} \cup Y_{\text{mis}} = \{y_i, r_i\}_{i=1}^n \). where \( Y_{\text{mis}} = \{r_i\}_{i=1}^n \) denote the missing data and \( R_1, \ldots, R_n \sim \text{TN}(0,1; (0, \infty)) \). From (10), the the surrogate function of \( \theta \) can be constructed as

\[ Q(\theta | \theta^{(t)}) = E[L(\theta | Y_{\text{com}})] | Y_{\text{obs}}, \theta^{(t)} = -n \log(\pi) - \frac{n}{2} \log(\sigma^2) + \frac{n}{2} \log(1 + \lambda^2) \]

\[ - \frac{1 + \lambda^2}{2\sigma^2} \sum_{i=1}^n \left[ (y_i - \mu_i)^2 - 2\sigma \lambda (y_i - \mu_i) r_i^{(t)} + \sigma^2 r_i^{(t)} \right] \]  \hspace{1cm} (13)

Where \( \theta^{(t)} \) denote the \( t \)-th approximation of the MLEs \( \hat{\theta} \)

\[ r_i^{(t)} | Y_{\text{obs}} \sim \text{TN} \left( \frac{\tau_i^{(t)}}{\sqrt{1 + \lambda_i^{(t)} + 1 + \lambda_i^{(t)} \Phi(\tau_i^{(t)})} \right) \]  \hspace{1cm} (14)

\[ r_i^{(t)} | Y_{\text{obs}} \sim \text{TN} \left( \frac{1}{1 + \lambda_i^{(t)}} + \frac{1}{1 + \lambda_i^{(t)} \Phi(\tau_i^{(t)})} \right) \]  \hspace{1cm} (15)

\[ \tau_i^{(t)} = \frac{\lambda (y_i - \mu_i^{(t)})}{\sigma \lambda^{(t)}}, \quad \mu_i^{(t)} = x_i \beta^{(t)} - \frac{\sigma^{(t)}}{3} \left[ m_0 (\lambda^{(t)}) + 2 \sqrt{\frac{2}{\pi}} \frac{\lambda^{(t)}}{\sqrt{1 + \lambda_{i}^{(t)}}} \right] \]  \hspace{1cm} (16)

Given an initial value \( \theta^{(0)} = \left( \beta^{(0)T}, \sigma^{(0)}, \lambda^{(0)} \right)^T \) for the median regression model. The E-Step of the EM algorithm is to calculate the conditional expectations (14)- (16). The M-Step is to update

\[ \theta^{(t+1)} = \theta^{(t)} + [-H(\theta^{(t)})]^{-1} G(\theta^{(t)}) \]  \hspace{1cm} (17)

where

\[ G(\theta^{(t)}) = \frac{\partial Q(\theta | \theta^{(t)})}{\partial \theta} = \left( \frac{\partial Q(\theta | \theta^{(t)})}{\partial \beta^T}, \frac{\partial Q(\theta | \theta^{(t)})}{\partial \sigma^2}, \frac{\partial Q(\theta | \theta^{(t)})}{\partial \lambda} \right) \]

\[ H(\theta^{(t)}) = \frac{\partial^2 Q(\theta | \theta^{(t)})}{\partial \theta \partial \sigma^T} \]  \hspace{1cm} (18)

### 4. Conclusion
We conduct Monte Carlo simulations to evaluate the finite sample performance of the proposed methods. The performance of the MLE $\hat{\theta}$ is assessed by the mean square error (MSE) defined by $\text{MSE}(\hat{\theta}) = E((\hat{\theta} - \theta)^T(\hat{\theta} - \theta))$. We next consider the median regression model (5), where the covariate vector $x_i = (x_{i1}, x_{i2}, x_{i3})^T$ with $x_{i1}, x_{i2}, x_{i3} \sim U(-1, 1)$. The values of parameters are set as $\beta = (4, -5, 6)^T$, $\sigma^2 = 2$, and $\lambda = 3, -3$. The sample size is set as $n = 100, 200, 300$. The results are shown in Table 1.

| $\lambda$ | model     | $\hat{\beta}^T$                  | $\hat{\sigma}^2$ | $\hat{\lambda}$ | $\text{MSE}(\hat{\beta}^T)$ | $\text{MSE}(\hat{\sigma}^2)$ | $\text{MSE}(\hat{\lambda})$ |
|----------|-----------|---------------------------------|------------------|-----------------|----------------------------|----------------------------|-----------------------------|
| 3        | 100       | (3.9987, -5.0007, 5.9975)       | 1.9441           | 2.8896          | 0.0117                   | 0.0142                     | 0.0397                      |
|          | 200       | (4.0004, -4.9987, 5.9999)       | 1.9730           | 2.9518          | 0.0055                   | 0.0064                     | 0.0165                      |
|          | 300       | (4.0007, -5.0013, 6.0007)       | 1.9851           | 2.9634          | 0.0037                   | 0.0043                     | 0.0115                      |
| -3       | 100       | (3.9985, -5.0011, 6.0026)       | 1.9460           | -2.8813         | 0.0119                   | 0.0149                     | 0.0419                      |
|          | 200       | (4.0013, -4.9987, 6.0010)       | 1.9737           | -2.9427         | 0.0056                   | 0.0061                     | 0.0179                      |
|          | 300       | (4.0025, -5.0000, 6.0005)       | 1.9843           | -2.9636         | 0.0035                   | 0.0040                     | 0.0107                      |

From Table 1, we have observed the following facts:

1. For a given sample size, the results for different values of $\lambda$ are similar in terms of MSE, showing that the performance of the parameter estimation method is not sensitive to the values of $\lambda$.

2. For a given $\lambda$, the MSEs of the MLEs $\hat{\beta}, \hat{\sigma}^2$ and $\hat{\lambda}$ become smaller and smaller as $n$ increases, indicating that the parameter estimation method performs well.

5. Application to Body Mass Index Data of Australia

In this section, we illustrate the proposed skew-normal median regression model by using the body mass index (BMI) data for 102 male and 100 female athletes collected at Australian Institute of Sport [14]. The BMI data consist of the response variable $Y$ (i.e., the BMI) in weight/height$^2$ and four covariates: $x_1$ (the red cell count), $x_2$ (the plasma ferritin concentration), $x_3$ (the body fat percentage) and $x_4$ (the lean body mass). We are interested in establishing the relationship between the BMI $Y$ and some important covariates.

The Figure 1 is histogram of $Y$, indicating that the distribution $Y$ approximately follows a skew-normal distribution. Therefore, based on the BMI data, we can consider the skew-normal median regression model. By applying the proposed EM algorithm in Section 3.2 to the skew-normal median regression model, we obtained the MLEs of parameters reported in the second column of Tables 2.

![Figure 1. The density estimate of $Y$, the body mass index.](image-url)
Table 2. MLEs of parameters for the median regression model and N regression model.

| Parameters | SN(median) | N(median/mean) |
|------------|------------|----------------|
| $\beta_0$  | 6.0690     | 4.6696         |
| $\beta_1$  | 0.4297     | 0.5320         |
| $\beta_2$  | 0.0072     | 0.0072         |
| $\beta_3$  | 0.2311     | 0.2550         |
| $\beta_4$  | 0.1703     | 0.1816         |
| $\sigma^2$ | 4.2190     | 0.0094         |
| $\lambda$  | 2.1683     | 0.0000         |

From Tables 2, it can be seen that the estimation results of multiple linear regression model and SN regression model are quite different. The skewness and kurtosis of BMI data of 202 athletes are 0.9465 and 5.1835 respectively. The kurtosis of the normal distribution is 3, however, the kurtosis of athletes is greater than 3 and the skewness is not 0. So there is a deviation in using normal distribution to fit the skew data. These results can provide valuable reference for the practical researchers.

6. Conclusion
The main aim of this paper is to establish the median regression model to capture the "medium" level of the skew-normal distribution. Our simulation studies and a real data analysis showed that the proposed model and method are useful in practice.

In this paper, we established the median regression model for the skew-normal distribution. A natural extension of the skew-normal distribution to other skewed distributions, e.g., skew-t-normal and skew-Laplace-normal distributions, may be one of our future works. Another future extension of this work is to consider the case of mixture regression model with multivariate outputs. Finally, one interesting direction is to extend the proposed models to partially linear models.

Acknowledgments
Liucang Wu's research was supported by a grant from the National Natural Science Foundation of China (No.11861041).

References
[1] Azzalini, A. A class of distributions which includes the normal ones[J]. Scandinavian Journal of Statistics, 1985, 12(2):171-178.
[2] Gomez, H.W., Venegas, O., Bolfarine, H. Skew-symmetric distributions generated by the distribution function of the normal distribution[J]. Environmetrics, 2007, 18(4): 395-407.
[3] Azzalini, A., Dalla, V.A. The multivariate skew-normal distribution[J]. Biometrika, 1996, 83(4): 715-726.
[4] Azzalini, A., Capitanio, A. Statistical applications of the multivariate skew normal distribution [J]. Journal of the Royal Statistical Society: Series B, 1999, 61(3):579-602.
[5] Azzalini, A., Capitanio, A. Distributions generate by perturbation of symmetry with emphasis on a multivariate skew-t distribution[J]. Journal of the Royal Statistical Society: Series B, 2003, 65(2):367-389.
[6] Cancho, V.G., Lachos, V.H., Ortega, E.M. A nonlinear regression model with skew-normal errors[J]. Statal Papers, 2010, 51(3):547-558.
[7] Vallea, R.B.A., Ozanb, S., Bolfarineb, C.H., Lachosb, V.H. Skew normal measurement error models[J]. Journal of Multivariate Analysis, 2005, 96:265-281.
[8] Kheradmandi, A., Rasekh, A. Estimation in skew-normal linear mixed measurement error models[J]. Journal of Multivariate Analysis, 2015, 136:1-11.
[9] Liu M., Lin T. A skew-normal mixture regression model[J]. Educational and Psychological Measurement, 2014, 74:139-162.
[10] Wu L., Zhang Z., Xu D. Variable selection in joint location and scale models of the skew-
normal distribution[J]. Journal of Statistical Computation & Simulation, 2013, 83(7):1266-1278.

[11] Li H., Wu L., Ma T. Variable selection in joint location, scale and skewness models based on the skew-normal distribution[J]. Journal of Systems Science and Complexity, 2017, 30(003): 694-709.

[12] Azzalini, A., Capitanio, A. The skew-normal and related families[M]. Cambridge University Press, New York, 2014.

[13] Dempster, A.P., Laird, N.M. Maximum likelihood from incomplete data via the EM algorithm (with discussions) [J]. Journal of the Royal Statistical Society: Series B, 1977, 39:1-38.

[14] Cook, R.D., Weisberg, S. An introduction to regression graphics[M]. John Wiley and Sons, New York, 1994.