Statistical measure researches for the impacts of COVID-19 on Shanghai Securities Composite Index: based on mode regression model using skew-normal distribution

Xin Zeng¹, Yuanyuan Ju¹, and Liucang Wu**

¹Faculty of Science, Kunming University of Science and Technology, Kunming, Yunnan Province, 650093, China
**Corresponding author’s e-mail: 11305029@kust.edu.cn

Abstract. The economy of China has been affected by the COVID-19 and it is necessary to measure and investigate it. The mode regression model and its data deletion model using skew-normal distribution are established via the industry sub-index data of Shanghai Stock Exchange. Statistical diagnosis method of the model is proposed by using Pena distance diagnosis statistics, and the unknown parameters are estimated by the EM algorithm facilitated by gradient descent (GDEM) method. The results show that the outliers are accurately identified by the proposed statistical diagnostic method, and parameter estimation results indicate that the contribution of all industries to economy during the COVID-19 is lower than before, among which industry and commerce are the hardest hit.

1. Introduction

Shanghai Stock Exchange takes industrial index, commercial index, real estate index, composite industry index and public industry index as the industry sub-index of China, which reflects the prosperity of each industry and the overall change of stock prices. The Shanghai Stock Composite Index (SSCI) is known as the "barometer" economic performance of China.

The mode of a distribution provides a significant summary when data involve asymmetric outcomes. The relevant research on identifying modes of population distributions can be found in [1-3]. A modal linear regression model was proposed by [4] via a nonparametric method. In the real world, most of the collected data are not strictly normal distribution and sometimes error structures in regression models no longer meet symmetric properties in many fields. In this case, it is inappropriate to use the normal distribution, t-distribution and other symmetric distributions to describe their properties. The skew-normal distribution, proposed by [5], is widely used to fit skewed data and good performance has been obtained. As an extension of the skew-normal distribution, the multivariate skew-normal distribution, skew-t distributions and multivariate skew-t distribution were developed by [6-8]. Based on the skew-normal distribution, [9] studied variable selection in joint location and scale models of the skew-normal distribution.

The mean modelling have been discussed extensively when the normality is assumed for the regression error in each cluster. Little work has been done for establishing mode regression models using skew-normal distribution (SNMRM) to study the effects of covariates on distribution characteristics. The SNMRM is potentially a very useful addition to current data analysis tools especially in the fields of economics, biomedicine. This also is the motivation of research conducted in the current article. Furthermore, in this paper, the statistical diagnosis problem of SNMRM is studied.
based on Pena distance diagnostic statistics and data deletion model. The proposed method is applied to the research for situation of economic development in the context of the COVID-19.

2. Skew-normal mode regression model

2.1. The skew-normal distribution

A random variable $Y$ is said to have a skew-normal distribution with location parameter $\mu$, scale parameter $\sigma \in (0, \infty)$ and skewness parameter $\lambda$, denoted by $Y \sim SN(\mu, \sigma^2, \lambda)$, if it has the probability density function

$$f_{SN}(y \mid \mu, \sigma^2, \lambda) = \frac{2}{\sigma} \phi\left(\frac{y - \mu}{\sigma}\right) \Phi\left(\frac{\lambda - \mu}{\sigma}\right),$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the probability density function and cumulative distribution function of the standard normal distribution, respectively. The stochastic representation for the random variable $Y$ can be given by

$$Y = \mu + \sigma \left[ \delta(\lambda)W + \frac{\delta(\lambda)V}{\lambda} \right],$$

where $\delta(\lambda) = \lambda / \sqrt{1 + \lambda^2}$, $V \sim N(0,1)$ and $W \sim TN(0,1)$, where $TN(\cdot)$ indicates a normal distribution truncated to $(0, \infty)$. The skew-normal distribution has hierarchical representation,

$$Y \mid (R = r) - N(\mu + r\delta(\lambda), \frac{\sigma^2}{1 + \lambda^2}) \text{ and } R \sim TN(0,1; (0, \infty))$$

[10] introduce the mode of the (1) can be expressed as

$$Mode(Y) = \mu + m_0(\lambda)\sigma$$

where

$$m_0(\lambda) \approx \mu_0(\lambda) - \frac{t_0(\lambda)\sigma_0(\lambda)}{2} - \frac{\text{sgn}(\lambda)}{2} \exp\left\{-\frac{2\pi}{|\lambda|}\right\},$$

and

$$\mu_0(\lambda) = \sqrt{\frac{2}{\pi}}\delta(\lambda), \quad \sigma_0^2(\lambda) = 1 - \mu_0^2(\lambda), \quad t_0(\lambda) = \frac{4 - \pi \mu_0^2(\lambda)}{2 \sigma_0^2(\lambda)},$$

where sign($\lambda$) indicates sign function for $\lambda$.

2.2. SNMRM

For the univariate case, let $Y_i \sim SN(\mu, \sigma^2, \lambda), i = 1, \cdots, n$. A linear mode regression model with skew-normal errors can be expressed as

$$y_i = x_i^T \beta + \varepsilon_i$$

where $Mode(Y) = x_i^T \beta$ and $Y_i$ is the response variable, $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ is the p-dimension vector of covariates, $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is an unknown p-dimension vector of the coefficients for mode regression, $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)$ is the vector of random errors and $\varepsilon_i \sim SN(-m_0(\lambda)\sigma, \sigma^2, \lambda)$. 

2.3. Data deletion model for SNMRM

We considered deleting the $i$-th observation, and judged whether the point was a strong influence point or an outlier by comparing the changes of Pena distance diagnostic statistics before and after deleting the point. The purpose of which is to study the impact of the $i$-th observation $(y_i, x_i)$ on the
overall model. The data deletion model for SNMRM is obtained after deleting the first data point of the original model and is given by

$$Y_{i(j)} \sim SN(\mu_{i(j)} , \sigma^2 , \lambda),$$

$$Mode(Y_{i(j)}) = x_i^T \beta_{i(j)} ,$$

where $i, j = 1, 2, \ldots, n, i \neq j$ , $\hat{\mu}_{i(j)}$ and $\hat{\beta}_{i(j)}$ represents the fitting value of the $i$ -th observation and estimated value of the regression coefficient after deleting the $j$ -th observation, respectively.

**Theorem 1.** The Pena distance diagnostic statistics of SNMRM is:

$$PD_i = \frac{1}{ph_{ij}} \sum_{j=1}^{n} h_{ij}^2 \hat{r}_{ij}^2 ,$$

where $\hat{r}_{ij} = \hat{\eta}_{ij} / (s_0 \sqrt{1-h_{ij}})$ is the $j$ -th standardized residual, $s_0$ and $\hat{\eta}_{ij} = y_{ij} - \hat{y}_{ij}$ indicate the standard deviation and residual, respectively.

**Proof.** The Pena distance of $i$ -th observation for SNMRM is defined as

$$S_i = \frac{s_i^T s_i}{p \cdot Var(\hat{y}_i)} ,$$

where $s_i = (\hat{y}_i - y_{i(1)} , \ldots , \hat{y}_i - y_{i(n)})$ and $\hat{y}_i - y_{i(j)} = h_{ij} \hat{\eta}_{ij} / (1-h_{ij})$ . $h_{ij} = x_i^T (X^T X)^{-1} x_i$ is the element in the $i$ -th row and $j$ -th column of the projection matrix $H$ . The diagonal element $h_{ij}$ is called the leverage value and the dimension of $H$ is denoted by $p$ . Thus,

$$s_i^T s_i = \sum_{j=1}^{n} \frac{h_{ij}^2 \hat{\eta}_{ij}^2}{(1-h_{ij})^2} ,$$

and $Var(\hat{y}_i) = x_i Var(\hat{\beta}) x_i^T = h_{ij} s_0^2$ , so we obtain

$$S_i = \frac{s_i^T s_i}{p \cdot Var(\hat{y}_i)} = \frac{1}{ph_{ij}} \sum_{j=1}^{n} \frac{h_{ij}^2 \hat{\eta}_{ij}^2}{1-h_{ij}} .$$

Theorem 1 is proved.

**Theorem 2.**

(a). When the sample contains no outliers, we have

$$E(S_i) \to \frac{1}{p} + \frac{(\mu_0 - m_0)^2}{\sigma^2} (n \to \infty) ,$$

(b). When the sample contains outliers, we have

$$E(S_i) \to 0$$

(c). We can get the same result as a normal distribution when $\lambda = 0$ , that is, $E(S_i) \to 1 / p$ .

The theorem 2 is obvious and we are not going to prove it.

3. EM algorithm

The joint probability density can be calculated from the hierarchical representation (3) of skew-normal distribution,

$$f_{Y|X}(y, r) = \frac{\lambda}{\pi \sigma \delta(\lambda)} \exp \left\{-\frac{1+\lambda^2}{2\sigma^2} \left[ e^2 - 2\sigma r e \delta(\lambda) + \sigma^2 r^2 \right] \right\} ,$$

where $e = y - \mu$ and $\mu = x^T \beta - m(\lambda) \sigma$ . According to Bayes criterion, we can obtain

$$R \mid (Y = y) \sim TN\left(\frac{e \delta(\lambda)}{\sigma} , \frac{1}{1+\lambda^2} ; (0, \infty) \right) ,$$

so
\[
E(R | Y = y) = \frac{e^{\delta(\lambda)}}{\sigma} + \frac{\delta(\lambda)}{\lambda} \phi(\tau), \tag{9}
\]

\[
E(R^2 | Y = y) = \frac{1}{1 + \lambda^2} + \frac{e^{\delta(\lambda)}}{\sigma} E(R | Y = y), \tag{10}
\]

where \( \tau = \lambda e/\sigma \).

Let \( R \) be the latent variable and form complete data. The complete-data log-likelihood function can be given by

\[
\ell(\theta | Y_{\text{con}}) = -n \log \pi - \frac{n}{2} \log(\sigma^2) - \frac{n}{2} \log(1 + \lambda^2) - \frac{1 + \lambda^2}{2\sigma^2} \left[ e_i^2 - 2\sigma e_i \delta(\lambda) r_i + \sigma^2 r_i^2 \right], \tag{11}
\]

where \( e_i = y_i - \mu_i \), \( \theta^{(t)} \) is denoted as parameter estimation for \( t \)-th iteration. The EM algorithm is as follows.

**E-Step.** The surrogate function can be calculated as

\[
Q(\theta | \theta^{(t)}) = E\left[ \ell(\theta | Y_{\text{con}}) | Y_{\text{obs}}, \theta^{(t)} \right] = -n \log(\pi) - \frac{n}{2} \log(\sigma^2) - \frac{n}{2} \log(1 + \lambda^2) - \frac{1 + \lambda^2}{2\sigma^2} \sum_{i=1}^{n} \left[ e_i^2 - 2\sigma e_i \delta(\lambda) r_i + \sigma^2 r_i^2 \right], \tag{12}
\]

where

\[
r_{1i}^{(t)} = E(R_1 | Y_{\text{obs}}, \theta^{(t)}) = \frac{(y_i - \mu_i^{(t)})\delta(\lambda^{(t)})}{\sigma^{(t)}} + \frac{\delta(\lambda^{(t)})}{\lambda^{(t)}} \phi(\tau_i^{(t)}) \Phi(\tau_i^{(t)}) \tag{13}
\]

\[
r_{2i}^{(t)} = E(R_2 | Y_{\text{obs}}, \theta^{(t)}) = \frac{1}{1 + \lambda^{(t)^2}} + \frac{(y_i - \mu_i^{(t)})\delta(\lambda^{(t)})}{\sigma^{(t)}} \phi(\tau_i^{(t)}) \Phi(\tau_i^{(t)}) \tag{14}
\]

and \( \tau_i^{(t)} = \lambda^{(t)} (y_i - \mu_i^{(t)})/\sigma^{(t)}, \quad \mu_i^{(t)} = x_i^T \beta^{(t)} - m_0(\lambda^{(t)}) \sigma^{(t)} \).

**M-Step.** Given two initial values \( \theta^{(0)} = (\beta^{(0)T}, \sigma^{(0)}, \lambda^{(0)}, \lambda^{(1)}) \), \( \theta^{(0)} = (\beta^{(0)T}, \sigma^{(0)}, \lambda^{(1)}) \). The parameter vector \( \theta^{(x+1)} \) was calculated via maximizing \( Q(\theta | \theta^{(t)}) \) with respect to \( \theta \) in M-Step. We calculate the MLEs of \( \theta \) by using the two-point step size gradient descent method[11] to update

\[
\theta^{(x+1)} = \theta^{(x)} + s^{(x)} G(\theta^{(x)}), \tag{15}
\]

where \( s^{(x)} \) is the step size defined by

\[
s^{(x)} = \frac{||[\theta^{(x)} - \theta^{(x-1)T}][G(\theta^{(x)}) - G(\theta^{(x-1)})]] ||}{||[G(\theta^{(x)}) - G(\theta^{(x-1)})]] ||}, \tag{16}
\]

and

\[
G(\theta^{(x)}) = \frac{\partial Q(\theta | \theta^{(x)})}{\partial \theta} = \begin{pmatrix}
\frac{\partial Q(\theta | \theta^{(x)})}{\partial \beta^T}, \\
\frac{\partial Q(\theta | \theta^{(x)})}{\partial \sigma^2}, \\
\frac{\partial Q(\theta | \theta^{(x)})}{\partial \lambda}
\end{pmatrix}^T, \tag{17}
\]

where

\[
\frac{\partial Q(\theta | \theta^{(x)})}{\partial \beta^T} = -\sum_{i=1}^{n} \frac{1 + \lambda^2}{\sigma^2} \left( e_i E_1 - \sigma r_i \delta(\lambda) E_1 \right),
\]

\[
\frac{\partial Q(\theta | \theta^{(x)})}{\partial \sigma^2} = \frac{1 + \lambda^2}{\sigma^2} \sum_{i=1}^{n} \left[ \frac{e_i^2}{\sigma} - e_i \delta(\lambda) r_i + e_i E_2 + \sigma r_i \delta(\lambda) E_2 \right],
\]

\[
- \frac{n}{\sigma},
\]
\[
\frac{\partial Q(\theta | \theta^{(i)})}{\partial \lambda} = -\sum_{i=1}^{n} \left[ \frac{e_i^2 \lambda}{\sigma^2} + \frac{e_i^2 (1 + \lambda^2)}{\sigma^2} E_3 - \frac{r_i^{(t)}}{\sigma} \left( \frac{e_i \delta(\lambda)}{\lambda} + 2 e_i \lambda \delta(\lambda) + \lambda^2 \delta'(\lambda) E_3 \right) + \lambda r_1^{(t)} \right] + \frac{n \delta^2(\lambda)}{\lambda},
\]
where \( E_1 = \frac{\partial e_i}{\partial \mu_i} = -x_i^T, E_2 = \frac{\partial e_i}{\partial \sigma} = m_0(\lambda), E_3 = \frac{\partial e_i}{\partial \lambda} = \sigma M, \) and
\[
M = \frac{\partial m_0(\lambda)}{\partial \lambda} = \frac{2}{\sqrt{\pi}} \frac{1}{(1 + \lambda^2)^{3/2}} \frac{\sigma_o(\lambda) T + t_0(\lambda) K}{2} \pi \cdot \text{sgn}(\lambda) \frac{2 \pi}{\lambda^2} e^{-\frac{2 \pi}{\lambda^2}},
\]
\[
T = \frac{\partial t_0(\lambda)}{\partial \lambda} = \frac{3(4 - \pi)}{2 \sigma_o^2(\lambda)} \left[ \frac{2}{\sqrt{\pi}} \mu_o^2(\lambda) \sigma_o(\lambda) \frac{\lambda^3}{(1 + \lambda^2)^{3/2}} - \mu_o^3(\lambda) S \right],
\]
\[
K = \frac{\partial \sigma_o(\lambda)}{\partial \lambda} = -\frac{2}{\sqrt{\pi}} \frac{\mu_o(\lambda)}{\sigma_o(\lambda)(1 + \lambda^2)^{3/2}}.
\]

The EM algorithm facilitated by the two-point step size gradient descent method in this part is used to iterate the parameters repeatedly until converges, and the MLEs of the parameters are obtained.

4. Application to a real data set
To demonstrate the performance of the proposed SNMRM and parameter estimation method. We consider a real data set about Shanghai Securities Composite Index (SSCI) and we are interested in the relationship between Shanghai Composite index \( y \) and the 5 performance measures given as the industrial index \( x_1 \), commercial index \( x_2 \), real estate index \( x_3 \), composite industry index \( x_4 \) and public industry index \( x_5 \). The data before (the Period 1, September 2, 2019 to October 31, 2019), during (the Period 2, February 3, 2020 to April 29, 2020) COVID-19 and work resumption (the Period 3, June 1, 2020 to August 31, 2020) are selected and note that the data are collected only on weekdays. In order to reduce the influence of heteroscedasticity on the estimation, a logarithmic approach is used. It is worth noting that the data of \( y \) from three periods is negatively skewed. The result can be found in the estimation of the skewness parameter of Table 1. Thus we say that \( Y \) approximates a skew-normal distribution, that is, \( Y_i \sim SN(\mu_i, \sigma^2, \lambda), \) where \( i = 1, \ldots, n_i \) and \( n_1, n_2, n_3 \) represent the sample sizes of three periods, respectively.

In order to compare the changes of \( y \) in three periods, we establish the following SNMRM. We
\[
\text{Mode}(Y_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4,
\]
Figure 1. Leverage value and distance Pena of SNMRM for SSCI data.

Table 1. Parameter estimation of SNMRM in three periods for COVID-19.

| Parameters | SNMRM          |
|------------|----------------|
|            | Period 1 | Period 2 | Period 3 |
| $\hat{\beta}_0$ | 0.0002   | 0.0007   | -0.0037  |
| $\hat{\beta}_1$ | 0.9100   | 0.4612   | 0.8387   |
| $\hat{\beta}_2$ | 0.8441   | 0.4004   | 0.7635   |
| $\hat{\beta}_3$ | 0.8759   | 0.4455   | 0.8087   |
| $\hat{\beta}_4$ | 0.8968   | 0.4869   | 0.8536   |
| $\hat{\sigma}^2$ | 1.0993   | 1.0995   | 0.0433   |
| $\hat{\lambda}$ | -0.4019  | -0.4016  | -0.4115  |

obtained the parameter estimation results of SNMRM in three different periods after removing outliers via data deletion model in Section 2.3. The results are shown in Table 1, where $\hat{\sigma}^2$ and $\hat{\lambda}$ are estimated values of scale parameters and skewness parameters respectively. From Table 1, we can clearly make the following conclusions: The $x_1, x_2, x_3, x_4$ and $x_5$ will all contribute to the growth of our economy, which is in line with expectations. However, we note that this contribution varies in size. The coefficients of all variables in period 1 are smaller than period 2, indicating that the contribution of all industries to economy in period 1 is lower. This reduction is almost half the size. In period 3, contribution level in all industries have basically returned to period 1.

5. Concluding remarks

In this article, the skew-normal distribution is used as regression error to overcome the potential weakness of normal distribution. Considering the asymmetric characteristics of skewed data, the main aim of this article is to consider a data diagnosis and parameter estimation method based on the skew-normal distribution for mode regression model (MRM). Traditional mean modeling is extended. The EM algorithm facilitated by the two-point step size gradient descent method (GDEM) is implemented in simulation studies and a real data set analysis. The GDEM keeps a faster convergence rate because it only uses the first derivative but the information of two step size is used in per iteration.

In this article, however, we only consider the MRM based on the skew-normal distribution. A natural extension of the skew-normal distribution is other skewed distributions, e.g. skew-t and skew-Laplace distributions. Furthermore, one interesting future direction is to extend the proposed models to
the finite mixture of models. Finally, the application of variable selection, semi-parametric and nonparametric methods for MRM may be one of our future works.

Acknowledgments
This work is partially supported by the National Natural Science Foundation of China under Grants 11861041, Academics and Technology Foundation of Kunming University of Science and Technology under Grants 2020YB208. The authors are grateful to anonymous referees for their careful corrections to and valuable comments on the original version of this article.

References
[1] Muller, D.W., Sawitzki, G. (1991) Excess mass estimates and tests for multimodality. J. Amer. Statist. Assoc., 86: 738–746.
[2] Davies, P.L., Kovac, A. (2004) Densities, spectral densities and modality. Ann. Statist., 32: 1093–1136.
[3] Yao, W.X., Lindsay, B.G. (2009) Bayesian mixture labeling by highest posterior density. J. Amer. Statist. Assoc., 104: 758–767.
[4] Yao, W.X., Li, L.H. (2014) A new regression model: modal linear regression, Scand. J. Stat., 41: 656–671.
[5] Azzalini, A. (1985) A class of distributions which includes the normal ones, Scand. J. Stat., 12: 171–178.
[6] Azzalini, A., Valle, A.D. (1996) The multivariate skew-normal distribution, Biometrika., 83: 715–726.
[7] Azzalini, A., Capitanio, A. (1999) Statistical applications of the multivariate skew normal distribution. J. R. Stat. Soc. Ser. B Stat. Methodol., 61: 579–602.
[8] Lin, T.I., Wang, W.L., McLachlan, G.J., Lee., S.X. (2018) Robust mixtures of factor analysis models using the restricted multivariate skew-t distribution. Statistical Modelling., 18: 50–72.
[9] Wu, L.C., Zhang, Z.Z., Xu, D.K. (2013) Variable selection in joint location and scale models of the skew-normal distribution. J. Stat. Comput. Simul., 83: 1266–1278.
[10] Azzalini, A., Capitanio, A. (2013) The Skew-Normal and Related Families. Cambridge University Press, New York.
[11] Barzilai, J., Borwein, J.M. (1988) Two-point step size gradient methods. IMA J. Numer. Anal., 8: 141–148.