ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 13, No. 1, 2018, *pp*.074-080

Local Influence Diagnostics of Replicated Data with Measurement Errors

Jingjing Lu, Hairong Li, Chunzheng Cao*
School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing 210044, China
(Received September 03, 2017, accepted December 20, 2017)

Abstract. Replicated data with measurement errors frequently exist in various scientific fields. In this work, we propose a replicated measurement error model for such data under scale mixtures of normal distributions. We consider local influence diagnostics to detect and classify outliers in the data through different perturbation schemes. A simulation study and an application confirm the effectiveness and robustness of the diagnostic method.

Keywords: scale mixtures of normal distributions, measurement error, local influence analysis, robustness, outliers

1. Introduction

Local influence analysis [1] is one of the effective ways to detect and classify outliers. Through various perturbation schemes on the established statistical model, it can detect the influential observations and make outlier discrimination. The latest research on this area can be seen, for example, in [2-5].

In this paper, we focus on outlier detection in replicated data with measurement errors. At first, we need to establish an appropriate model to depict the correlation between repeated measurements data as well as to characterize the effect of measurement errors on the data. Generally, the model is based on the assumption of normal distribution [6,7]. Recently, Cao et al. [8,9] proposed a replicated measurement error model under heavy-tailed distribution, which can bring us robust inferences. In this paper, we study local influence analysis on the heavy-tailed replicated measurement error model under different perturbation schemes. We aim to achieve an effective and robust diagnostic method for outlier detection in replicated measurement data.

The paper is organized as follows. In Section 2, we give the diagnostic methodology, including a brief description of the heavy-tailed replicated measurement error model and the local influence approach. In Section 3, we carry out numerical simulation. In Section 4, we display an application on a real data. We give a brief conclusion in the last section.

2. Methodology

2.1 The model

Let ξ_t and η_t $(t=1,\ldots,n)$ represent the true values of the explanatory variable and the response variable in the observations respectively. Their corresponding actual repeated measurement data are $x_t^{(i)}, i=1,\cdots,p$ and $y_t^{(j)}, j=1,\cdots,q$, which satisfy a replicated measurement error model

$$\begin{aligned} x_t^{(i)} &= \xi_t + \delta_t^{(i)}, \ \ i=1,\ldots,p, \\ y_t^{(j)} &= \eta_t + \varepsilon_t^{(j)}, \ \ j=1,\ldots,q, \\ \eta_t &= \alpha + \beta\xi_t, \ \ t=1,\ldots,n, \end{aligned} \tag{1}$$

Where δ and ε are measurement errors. Let $\boldsymbol{Z}_t = (x_t^{(1)},\ldots,x_t^{(p)}, y_t^{(1)},\ldots,y_t^{(q)})^T$ be the actual observations. Unlike the traditional normality assumption, here we propose a hierarchical distribution structure for $\boldsymbol{Z}_t$:

* Corresponding author. *E-mail address*: caochunzheng@163.com.

$$\begin{aligned} \boldsymbol{Z}_t \mid \xi_t, \upsilon_t &\overset{ind}{\sim} N_m(\boldsymbol{a}+\boldsymbol{b}\xi_t, \kappa(v_t)\mathbf{D}(\boldsymbol{\phi})), \\ \xi_t \mid \upsilon_t &\overset{ind}{\sim} N(\lambda, \kappa(v_t)\varphi_\xi),\ v_t \overset{iid}{\sim} H(v;\upsilon), t=1,\ldots,n, \end{aligned} \tag{2}$$

where $m=p+q$, $\boldsymbol{a}=(0,\ldots,0,\alpha\mathbf{1}_q^T)^T$, $\boldsymbol{b}=(\mathbf{1}_p^T,\beta\mathbf{1}_q^T)^T$, $\mathbf{1}_p$ and $\mathbf{1}_q$ represent p-dimensional and q-dimensional vector of ones respectively, $\boldsymbol{\phi}=(\varphi_\delta\mathbf{1}_p^T,\varphi_\varepsilon\mathbf{1}_q^T)^T$, $\mathbf{D}(\cdot)$ denotes the diagonal transformation that transforms a vector to a diagonal matrix. The latent variable v_t can adjust the weight of the influence of different samples on the parameter estimation, so as to obtain a robust inference effect. Statistical inference of this model can be found in [8].

2.2 The local influence approach

The purpose of local influence is to summarize the behavior of some influence measure $T(\boldsymbol{\omega})$ when small perturbations take place in the data or model, where $\boldsymbol{\omega}=(\omega_1,\cdots,\omega_g)$ is a g-dimensional perturbation vector. Let $\boldsymbol{\theta}=(\lambda,\alpha,\beta,\varphi_\delta,\varphi_\varepsilon,\varphi_\xi)^T$ be the parameter vector of model (1), $\boldsymbol{Z}_c=(\boldsymbol{Z},\boldsymbol{\xi},v)$ be the complete-data, where $\boldsymbol{Z}=(\boldsymbol{Z}_1^T,\ldots,\boldsymbol{Z}_n^T)^T$, $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_n)^T$, $\boldsymbol{v}=(v_1,\ldots,v_n)^T$, $\hat{\kappa}_t=\mathrm{E}[\kappa^{-1}(v_t)\mid\hat{\boldsymbol{\theta}},\boldsymbol{Z}_t]$, $\hat{\tau}=\hat{\varphi}_\xi/(1+\hat{\varphi}_\xi\hat{\mathbf{b}}^T\mathbf{D}^{-1}(\hat{\boldsymbol{\phi}})\hat{\mathbf{b}})$. Let $l_c(\boldsymbol{\theta},\boldsymbol{\omega}|\boldsymbol{Z}_c)$ be the log-likelihood of the perturbed model for the complete-data. We assume that there is an $\boldsymbol{\omega}_0$ such that $l_c(\boldsymbol{\theta},\boldsymbol{\omega}_0\mid\mathbf{Z}_c)=l_c(\boldsymbol{\theta}\mid\mathbf{Z}_c)$ for all $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}(\boldsymbol{\omega})$ be the maximum likelihood estimation of $\boldsymbol{\theta}$ under the function $Q(\boldsymbol{\theta},\boldsymbol{\omega}\mid\hat{\boldsymbol{\theta}})=\mathrm{E}\{l_c(\boldsymbol{\theta},\boldsymbol{\omega}\mid\boldsymbol{Z}_c)\mid\hat{\boldsymbol{\theta}},\boldsymbol{Z}\}$. We define $\boldsymbol{\alpha}(\boldsymbol{\omega})=(\boldsymbol{\omega}^T,f_Q(\boldsymbol{\omega}))^T$ as the influence graph, where $f_Q(\boldsymbol{\omega})=2\{Q(\hat{\boldsymbol{\theta}}\mid\hat{\boldsymbol{\theta}})-Q(\hat{\boldsymbol{\theta}}(\boldsymbol{\omega})\mid\hat{\boldsymbol{\theta}})\}$ is the Q-displacement function which can describe the difference between $\hat{\boldsymbol{\theta}}(\boldsymbol{\omega})$ and $\hat{\boldsymbol{\theta}}$. The curvature $C_{f_Q,\mathbf{d}}=-2\mathbf{d}^T\ddot{\mathbf{Q}}_{\boldsymbol{\omega}_0}\mathbf{d}$ of $\boldsymbol{\alpha}(\boldsymbol{\omega})$ in the direction of the unit vector $\mathbf{d}$ at $\boldsymbol{\omega}_0$ can investigate the behavior of the Q-displacement function, where $-\ddot{\mathbf{Q}}_{\boldsymbol{\omega}_0}=\boldsymbol{\Delta}_{\boldsymbol{\omega}_0}^{\mathbf{T}}\{-\ddot{\mathbf{Q}}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})\}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\omega}_0}$, $\boldsymbol{\Delta}_{\boldsymbol{\omega}_0}$ is the value of $\boldsymbol{\Delta}_{\boldsymbol{\omega}}=\partial Q(\boldsymbol{\theta},\boldsymbol{\omega}\mid\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\omega}^T$ without disturbance. The Hessian matrix $\ddot{\mathbf{Q}}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})=\partial^2Q(\boldsymbol{\theta}\mid\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathbf{T}}$ has elements given by (the elements not shown are all 0. The following is the same)

$$\ddot{Q}_{\lambda\lambda}=-\frac{1}{\varphi_\xi}\sum_{t=1}^n\hat{\kappa}_t,\ \ddot{Q}_{\lambda\varphi_\xi}=-\frac{1}{\varphi_\xi^2}\sum_{t=1}^n(\hat{\kappa}_t(\hat{\xi}_t-\lambda)),\ \ddot{Q}_{\alpha\alpha}=-\frac{q}{\varphi_\varepsilon}\sum_{t=1}^n\hat{\kappa}_t,\ \ddot{Q}_{\alpha\beta}=-\frac{q}{\varphi_\varepsilon}\sum_{t=1}^n\hat{\kappa}_t\hat{\xi}_t,$$

$$\ddot{Q}_{\alpha\varphi_\varepsilon}=-\frac{1}{\varphi_\varepsilon^2}\sum_{t=1}^n[\hat{\kappa}_t\sum_{j=1}^q(y_{tj}-\alpha-\beta\hat{\xi}_t)],\ \ddot{Q}_{\beta\beta}=-\frac{q}{\varphi_\varepsilon}\sum_{t=1}^n\hat{\kappa}_t\hat{\xi}_t^2-\frac{q}{\varphi_\varepsilon}\hat{\tau},$$

$$\ddot{Q}_{\beta\varphi_\varepsilon}=-\frac{1}{\varphi_\varepsilon^2}\sum_{t=1}^n[\hat{\kappa}_t\hat{\xi}_t\sum_{j=1}^q(y_{tj}-\alpha-\beta\hat{\xi}_t)]+\frac{\hat{\tau}}{\varphi_\varepsilon^2}q\beta,\ \ddot{Q}_{\varphi_\xi\varphi_\xi}=\frac{1}{2\varphi_\xi^2}-\frac{1}{\varphi_\xi^3}\sum_{t=1}^n[\hat{\kappa}_t(\hat{\xi}_t-\lambda)^2+\hat{\tau}],$$

$$\ddot{Q}_{\varphi_\delta\varphi_\delta}=\frac{p}{2\varphi_\delta^2}-\frac{1}{\varphi_\delta^3}\sum_{t=1}^n[\hat{\kappa}_t\sum_{i=1}^p(x_{ti}-\hat{\xi}_t^2)^2]-\frac{\hat{\tau}p}{\varphi_\delta^3},\ \ddot{Q}_{\varphi_\varepsilon\varphi_\varepsilon}=\frac{q}{2\varphi_\varepsilon^2}-\frac{1}{\varphi_\varepsilon^3}\sum_{t=1}^n[\hat{\kappa}_t\sum_{j=1}^q(y_{tj}-\alpha-\beta\hat{\xi}_t)^2]-\frac{\hat{\tau}}{\varphi_\varepsilon^3}q\beta^2.$$

In this section we consider three different perturbation schemes: case-weight perturbation, response variable perturbation and variance ratio perturbation. The key step is to calculate the elements of the matrix $\boldsymbol{\Delta}_{\boldsymbol{\omega}_0}$.

i) Case-weight perturbation

We consider an arbitrary attribution of weights for the expected complete-data log-likelihood function called perturbed Q-function, which is presented by $Q(\boldsymbol{\theta},\boldsymbol{\omega}\mid\hat{\boldsymbol{\theta}})=\sum_{t=1}^n\omega_t\mathrm{E}[l_{c,t}(\boldsymbol{\theta}\mid\mathbf{Z}_{c,t})\mid\hat{\boldsymbol{\theta}},\mathbf{Z}_t]$, where $\boldsymbol{\omega}=(\omega_1,\ldots,\omega_n)^T$ is an $n\times1$ vector with $\boldsymbol{\omega}_0=(1,\ldots,1)^T$. The matrix $\boldsymbol{\Delta}_{\boldsymbol{\omega}_0}$ has elements given by

$$\frac{\partial^2Q(\boldsymbol{\theta}\mid\hat{\boldsymbol{\theta}})}{\partial\lambda\partial\omega_t}|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}=\frac{1}{\varphi_\xi}\sum_{t=1}^n[\hat{\kappa}_t(\hat{\xi}_t-\lambda)],\ \frac{\partial^2Q(\boldsymbol{\theta}\mid\hat{\boldsymbol{\theta}})}{\partial\alpha\partial\omega_t}|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}=\frac{1}{\varphi_\varepsilon}\sum_{t=1}^n[\hat{\kappa}_t\sum_{j=1}^q(y_{tj}-\alpha-\beta\hat{\xi}_t)],$$

$$\frac{\partial^2Q(\boldsymbol{\theta}\mid\hat{\boldsymbol{\theta}})}{\partial\beta\partial\omega_t}|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}=\frac{1}{\varphi_\varepsilon}\sum_{t=1}^n[\hat{\kappa}_t\hat{\xi}_t\sum_{j=1}^q(y_{tj}-\alpha-\beta\hat{\xi}_t)]-\frac{\hat{\tau}}{\varphi_\varepsilon}q\beta,$$

$$\frac{\partial^2Q(\boldsymbol{\theta}\mid\hat{\boldsymbol{\theta}})}{\partial\varphi_\xi\partial\omega_t}|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}=-\frac{1}{2\varphi_\xi}+\frac{1}{2\varphi_\xi^2}\sum_{t=1}^n[\hat{\kappa}_t(\hat{\xi}_t-\lambda)^2+\tau^2],$$

$$\frac{\partial^2 Q(\boldsymbol{\theta}\mid\hat{\boldsymbol{\theta}})}{\partial\varphi_\delta\partial\omega_t}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}=-\frac{p}{2\varphi_\delta}+\frac{1}{2\varphi_\delta^2}\sum_{t=1}^{n}[\hat{\kappa}_t\sum_{i=1}^{p}(x_t^{(i)}-\hat{\xi}_t)^2]+\frac{\hat{\tau}p}{2\varphi_\delta^2},$$

$$\frac{\partial^2 Q(\boldsymbol{\theta}\mid\hat{\boldsymbol{\theta}})}{\partial\varphi_\varepsilon\partial\omega_t}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}=-\frac{q}{2\varphi_\varepsilon}+\frac{1}{2\varphi_\varepsilon^2}\sum_{t=1}^{n}[\hat{\kappa}_t\sum_{j=1}^{q}(y_t^{(j)}-\alpha-\beta\hat{\xi}_t)^2]+\frac{\hat{\tau}q\beta^2}{2\varphi_\varepsilon^2}.$$

ii) Response variable perturbation

This perturbation scheme, which can diagnose observations with great influence on their prediction, is represented by replacing $y_t^{(j)}$ with $y_{t\omega}^{(j)}=y_t^{(j)}+\omega_t$, with $\boldsymbol{\omega}_0=(0,\cdots,0)^T$ meaning non-perturbed model. In this scheme, the matrix $\boldsymbol{\Delta}_{\boldsymbol{\omega}_0}$ is given by

$$\frac{\partial^2 Q(\boldsymbol{\theta}\mid\hat{\boldsymbol{\theta}})}{\partial\alpha\partial\omega_t}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}=\frac{q}{\varphi_\varepsilon}\sum_{t=1}^{n}\hat{\kappa}_t,\ \frac{\partial^2 Q(\boldsymbol{\theta}\mid\hat{\boldsymbol{\theta}})}{\partial\beta\partial\omega_t}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}=\frac{q}{\varphi_\varepsilon}\sum_{t=1}^{n}(\hat{\kappa}_t\hat{\xi}_t),\ \frac{\partial^2 Q(\boldsymbol{\theta}\mid\hat{\boldsymbol{\theta}})}{\partial\varphi_\varepsilon\partial\omega_t}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}=\frac{1}{\varphi_\varepsilon^2}\sum_{t=1}^{n}(\hat{\kappa}_t\sum_{j=1}^{q}(y_{tj}-\alpha-\beta\hat{\xi}_t)).$$

iii) Variance ratio perturbation

This perturbation is introduced by writing $\varphi_{\varepsilon t,\boldsymbol{\omega}}=\omega_t\varphi_\varepsilon$. In this case, the matrix $\boldsymbol{\Delta}_{\boldsymbol{\omega}_0}$, the non-perturbed case can be acquired when $\boldsymbol{\omega}_0=(1,\ldots,1)^T$. The matrix $\boldsymbol{\Delta}_{\boldsymbol{\omega}_0}$ is given by

$$\frac{\partial^2 Q(\boldsymbol{\theta}\mid\hat{\boldsymbol{\theta}})}{\partial\alpha\partial\omega_t}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}=-\frac{1}{\varphi_\varepsilon}\sum_{t=1}^{n}[\hat{\kappa}_t\sum_{j=1}^{q}(y_{tj}-\alpha-\beta\hat{\xi}_t)],$$

$$\frac{\partial^2 Q(\boldsymbol{\theta}\mid\hat{\boldsymbol{\theta}})}{\partial\beta\partial\omega_t}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}=-\frac{1}{\varphi_\varepsilon}\sum_{t=1}^{n}[\hat{\kappa}_t\hat{\xi}_t\sum_{j=1}^{q}(y_{tj}-\alpha-\beta\hat{\xi}_t)]+\frac{\hat{\tau}}{\varphi_\varepsilon}q\beta,$$

$$\frac{\partial^2 Q(\boldsymbol{\theta}\mid\hat{\boldsymbol{\theta}})}{\partial\varphi_\varepsilon\partial\omega_t}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}=\frac{q}{2\varphi_\varepsilon}-\frac{1}{\varphi_\varepsilon^2}\sum_{t=1}^{n}(\hat{\kappa}_t\sum_{j=1}^{q}(y_{tj}-\alpha-\beta\hat{\xi}_t)^2)-\frac{1}{\varphi_\varepsilon^2}\hat{\tau}q\beta^2.$$

3. Simulation study

We conduct numerical studies to evaluate the performance of several diagnostic methods. The proposed local influence will be compared with the method based on adjusted Pearson residuals. To generate data of model (1), we consider two types of distribution: normal distribution (N) and student-*t* distribution (T). The true values of the parameters are set as follows: $\lambda=2$, $\alpha=3$, $\beta=1$, $\phi_\delta=1$, $\phi_\varepsilon=0.5$, $\phi_\xi=1$. The number of repeated measurements is $p=3$, $q=2$. Here, we choose the response variable perturbation as an example. In order to add an outlier, we shift the value of the 12-th sample x_{12} to $x_{12}^*=x_{12}+\omega\mu$. The parameter ω=2 represents the degree of perturbation. We will calculate type I error and type II error to compare the efficiency of the two diagnostic methods. The hypothesis test can be expressed as

$$H_1\text{: The 12-th sample is a normal one} \leftrightarrow H_2\text{: The 12-th sample is an outlier.} \quad (3)$$

Here, type I error means the probability of the 12th sample is the normal one but is classified as the outlier, and type II error is the probability of the 12-th sample is the outlier but is classified as the normal one.

Firstly, the data is generated from model (1) under the normal distribution. We do $N=1000$ replications with sample size $n=30$. Table 1 reports the results of type I and II errors based on the sampling data using the two diagnostic methods. For local influence, type I and type II errors are both ideally small ($I_{loc\_\inf}=0.0030$, $II_{loc\_\inf}=0$); while for the adjusted Pearson residual analysis, type I error is very high ($I_{res}=0.8470$) to keep the type II error as 0. The results indicate that local influence is more effective than the residual based method when the data comes from normal distribution.

Table 1: The test results of two diagnostic methods under the normal distribution

Model	Local influence		Adjusted Pearson residual	
	Type I error	Type II error	Type I error	Type II error
N	0.0030	0	0.8470	0

Secondly, we generate data from model (1) under the *t* distribution. All the designed values of this scenario are the same as before. The test results are listed in Table 2. Since statistical inference based on heavy-tailed model will be more robust than the normal model, the influence of the outliers on the diagnostic result is reasonably slight. This is the reason why the type II error under local influence is larger than it under residual method. However, the type I error under local influence is still smaller than it under residual method. Overall,

the local influence analysis based on the heavy-tailed distribution is more effective and more robust than the normal one.

Table 2: The test results of two diagnostic methods under the t distribution

Model	Local influence		Adjusted Pearson residual	
	Type I error	Type II error	Type I error	Type II error
t	0	1	0.8900	0

4. Application

In this section, we consider the CSFII data[10] as an illustrative example. The data contains the 24-h recall measures and three additional 24-h recall phone interviews of 1,722 women's diet habits. We consider the calorie intake/5,000 as ξ, and the saturated fat intake/100 as η. Instead of ξ and η, the nutrition variables x and y are recorded by four 24-h recalls, which are assumed to follow model (1) with $p=q=4$. We consider four distributions for comparison, which include the normal distribution (N), the t distribution (T), the slash distribution (SL) and the contaminated normal distribution (CN).

4.1. Local influence

According to the spectral analysis, we have $-2\ddot{\mathbf{Q}}_{\omega_0}=\sum_{k=1}^{g}\lambda_k\mathbf{e}_k\mathbf{e}_k^T$, where $(\lambda_1,\mathbf{e}_1),\ldots,(\lambda_g,\mathbf{e}_g)$ are the eigenvalues-eigenvectors with eigenvalues $\lambda_1\geq\ldots\geq\lambda_h$, $\lambda_{h+1}=\ldots=\lambda_g=0$ and orthogonal eigenvectors $\mathbf{e}_1,\ldots,\mathbf{e}_g$. Let $\tilde{\lambda}_k=\lambda_k/(\lambda_1+\ldots+\lambda_h)$, $\mathbf{e}_k^2=(e_{k1}^2,\ldots,e_{kg}^2)$, $M(0)=\sum_{k=1}^{h}\tilde{\lambda}_k\mathbf{e}_k^2$. Let $\bar{M}(0)$ and $SM(0)$ denote, respectively, the mean and standard deviation of $\{M(0)_l, l=1,\cdots,g\}$. Observations with value of $M(0)_l$ significantly greater than $\bar{M}(0)+c^*SM(0)$ ($c^*=4$) are considered as potential outliers[11].

i) Case-weight perturbation

The local influence intensity of each observations under this perturbation are plotted in Figure 1. The numbers of outliers under different distributions are listed in Table 3.

The values of baselines are calculated based on the variance of the intensity. It is not difficult to find that the intensities of outliers under the normal distribution is much higher than those under the heavy-tailed distributions. For example, the intensity of the strongest influential 1421 under the normal distribution is greater than 0.04, while the intensity of the strongest influential 1569 under T and SL is less than 0.018. From this viewpoint, the diagnoses under the heavy-tailed distributions are much slightly affected by the outliers than those under the normal distribution.

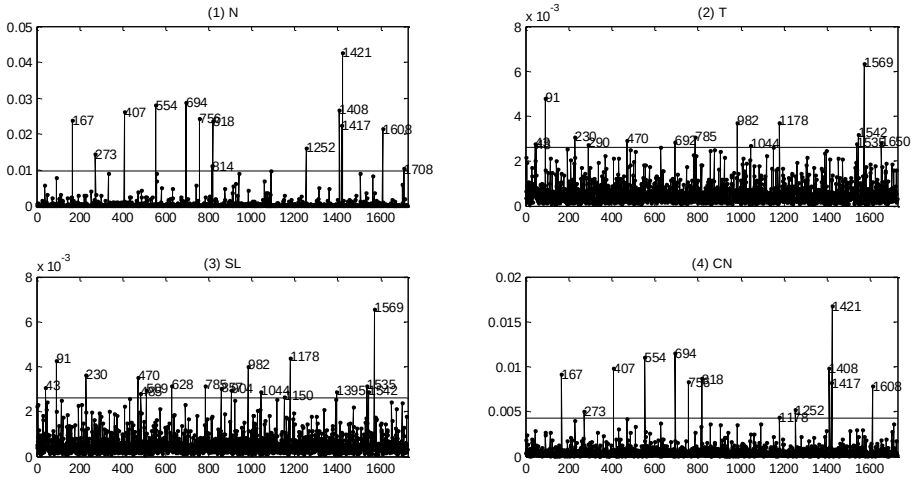

Fig. 1: Local influence under case-weight perturbation

Table 3: Outlier information under case-weight perturbation

Model	The number of outliers
N	167,273,403,554,694,756,814,818,1252,1408,1417,1421,1608,1708
T	43,48,91,230,290,470,692,785,985,1044,1178,1535,1542,1569,1650
SL	43,91,230,470,485,509,628,785,857,904,982,1044,1178,1395,1535,1543,1569
CN	167,273,407,554,694,756,818,1178,1252,1421,1408,1417,1608

ii) Response variable perturbation

Figure 2 shows the influence intensity under different distributions. Under the normal distribution, the intensity of the strongest influential 1252 is about 0.005. But the strongest influential under T and SL are 370 and 982 respectively, with intensity about 0.0025. The strongest influential 1178 under CN is about 0.003. Furthermore, the number of potential outliers are listed in Table 4. We find that both the number and the intensity of the influential observations are clearly reduced when we use heavy-tailed distributions. This results coincide with the conclusion drawn from case-weight perturbation.

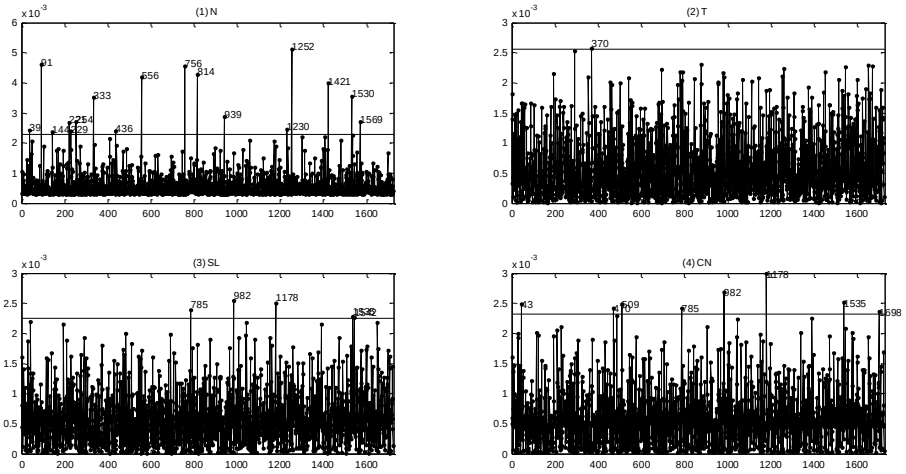

Fig. 2: Local influence under response variable perturbation

Table 4: Outlier information under response variable perturbation

Model	The number of outliers
N	39,91,144,221,229,254,333,436,556,756,814,939,1230,1252,1421,1530,1569
T	370
SL	785,982,1178,1535,1542
CN	43,470,509,785,982,1178,1535,1698

iii) Variances ratio perturbation

Figure 3 and Table 5 give respectively, the intensity of the influential and the information of the outliers under this perturbation. The diagnostic results under N and CN are similar. However, they are obviously different from the other two heavy-tailed distributions. For example, under the normal distribution, the intensity of the strongest influential 1421 is about 0.07, while under *t* distribution the strongest influential 1090 is only 0.005, and the strongest influential 75 under the slash distribution is less than 0.004. These results help us have a more objective understanding of the outliers and avoid the misdiagnosis caused by misclassification of the distribution.

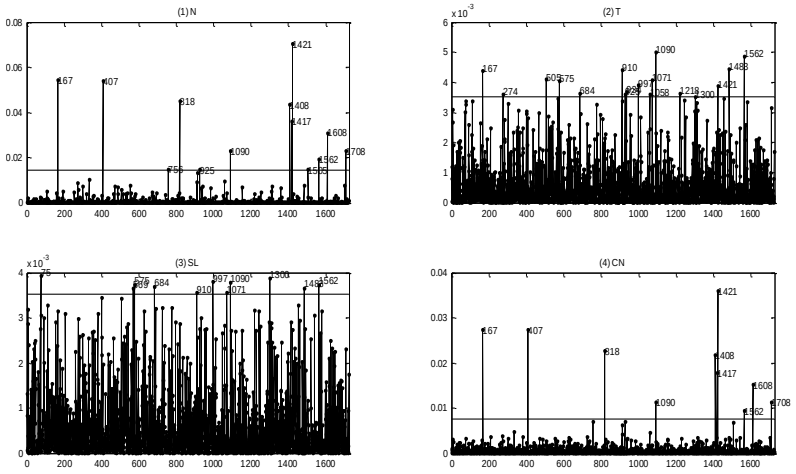

Fig. 3: Local influence under variances ratio perturbation

Table 5: Outlier information under variances ratio perturbation

Model	The number of outliers
N	167,407,756,818,925,1090,1408,1417,1421,1505,1562,1608,1708
T	167,274,505,575,684,910,925,934,997,1058,1071,1090,1218,1300,1421,1483,1562
SL	75,569,575,684,910,997,1071,1090,1300,1483,1562
CN	167,407,818,1090,1408,1417,1421,1562,1608,1708

4.2. Global influence

Through the local influences above, we treat the following eight cases as the most potential outliers: 1421, 1569, 1252, 370, 982, 1178, 1090 and 75. In order to reveal the impact of these eight observations on the parameter estimation, we remove them to obtain the maximum likelihood estimation $\hat{\boldsymbol{\theta}}^0$. Lu and Song[12] suggest that the following two quantities can measure the difference between the original maximum likelihood estimation and $\hat{\boldsymbol{\theta}}^0$

$$TRC = \sum_{j=1}^{n_p} |(\hat{\theta}_j - \hat{\theta}_j^0)/\hat{\theta}_j|, \ MRC = \max_{j=1,\cdots,n_p} |(\hat{\theta}_j - \hat{\theta}_j^0)/\hat{\theta}_j|,$$

where n_p is the number of parameters. The scale parameters in different distributions cannot be compared directly, so we calculate TRC and MRC only for the location parameters and list the results in Table 6. We can discover that the greatest changes take place under the normal distribution. The TRC and MRC are the smallest under T. The diagnostic similarity between SL and T shows that the heavy-tailed distribution not only has a good robustness to outliers, but also can overcome misspecification of distribution to some extent.

Table 6: TRC and MRC of global influence under the four distributions

Model	TRC	MRC
N	0.1732	0.1611
T	0.0476	0.0436
SL	0.0518	0.0479
CN	0.0729	0.0673

5. Conclusion

The detection and classification of outliers play important roles in data mining. In this work, we construct diagnostics of local influences based on a proposed replicated measurement error model. Numerical analysis confirms the effect of diagnostic methods under the heavy-tailed distributions. By comparing the results of different heavy-tailed distributions, we can obtain a more objective understanding of the data. The method proposed in this work can be extended to a wider range of use.

6. References

[1] R. D. Cook. Assessment of local influence (with discussion). Journal of the Royal Statistical Society Series B, 1986,

48(2): 133-169.

[2] V. H. Lachos, T. Angolini, C. A. Abanto-Valle. On estimation and local influence analysis for measurement errors models under heavy-tailed distributions. Statistical Papers, 2011, 52(3): 569-590.

[3] C. Z. Cao, J. G. Lin, J. Q. Shi. Diagnostics on nonlinear model with scale mixtures of skew-normal and first-order autoregressive errors. Statistics, 2014, 48(5): 1033-1047.

[4] Q. Gao, M. Ahn, H. Zhu. Cook's distance measures for varying coefficient models with functional responses. Technometrics, 2015, 57(2): 268-280.

[5] T. W. Rakhmawati, G. Molenberghs, G. Verbeke, C. Faes. Local influence diagnostics for generalized linear mixed models with overdispersion. Journal of Applied Statistics, 2017, 44(4): 620-641.

[6] Y. Isogawa. Estimating a multivariate linear structural relationship with replication. Journal of the Royal Statistical Society Series B, 1985, 47(2): 211-215.

[7] N. Lin, B. A. Bailey, X. M. He, W. G. Buttlar. Adjustment of measuring devices with linear model. Technometrics, 2004, 46(2): 127-134.

[8] J. G. Lin, C. Z. Cao. On estimation of measurement error models with replication under heavy-tailed distributions. Computational Statistics, 2013, 28(2): 809-829.

[9] C. Z. Cao, J. G. Lin, J. Q. Shi, X. Zhang. Multivariate measurement error models for replicated data under heavy-tailed distributions. Journal of Chemometrics, 2015, 29(8): 457-466.

[10] F. E. Thompson, M. F. Sowers, F. E. Jr, B. J. Parpia. Sources of fiber and fat in diets of US women aged 19 to 50: implications for nutrition education and policy. American Journal of Public Health, 1992, 82(5): 695-702.

[11] H. Zhu, L. Sik-Yum. Local influence for incomplete data models. Journal of the Royal Statistical Society Series B, 2001, 63(1): 111-126.

[12] B. Lu, X. Y. Song. Local influence analysis of multivariate probit latent variable models. Journal of Multivariate Analysis, 2006, 97(8): 1783-1798.