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

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

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

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