COVRATIO statistic for replicated linear functional relationship model

A Mohd Arif\(^1,2\), Y Z Zubairi\(^3\) and A G Hussin\(^4\)

\(^1\) Institute for Advanced Studies, University of Malaya, 50603 Kuala Lumpur, Malaysia
\(^2\) Centre for Foundation Studies, National Defence University of Malaysia, 57000 Kuala Lumpur, Malaysia
\(^3\) Centre for Foundation Studies in Science, University of Malaya, 50603 Kuala Lumpur, Malaysia
\(^4\) Faculty of Defence Science and Technology, National Defence University of Malaysia, 57000 Kuala Lumpur, Malaysia

E-mail: yzulina@um.edu.my

Abstract. Replicated linear functional relationship model (LFRM) can be categorized under the errors-in-variables model where variables involved are measured with error. However, the presence of outliers in dataset significantly impacts the parameter estimation. We extend the use of the COVRATIO statistic which has been successfully used in unrepli- cated LFRM for detecting the outliers. A simulation study is used to obtain the cut-off point at 10\% upper percentiles. An illustration of this procedure is presented for its potential in a real data set. The procedure successfully identifies the outlier present in the data set.

COVRATIO statistic; outlier detection; replicated linear functional relationship model

1. Introduction

A linear functional relationship model (LFRM) is one of the branch under the errors-in-variables model including structural relationship model and ultrastructural relationship model. Furthermore, linear functional relationship model can be divided into unreplicated and replicated LFRM with certain recommendations [1]. In a replicated linear functional relationship model, the assumption on the ratio of error variances is no longer needed unlike the unreplicated LFRM [2]. Over the years, many authors have been discussed about parameter estimation for replicated model [3-7].

In replicated LFRM, there exists a linear relationship between \( X_i \) and \( Y_i \) as follows:

\[
Y_i = \alpha + \beta X_i \quad \text{for} \quad i = 1, 2, \ldots, p
\]  

where \( \alpha \) is the intercept and \( \beta \) is the slope parameters respectively. In this model, both \( X_i \) and \( Y_i \) variables are subject to errors \( \delta_{ij} \) and \( \epsilon_{ik} \) respectively. The equation can be written as:

\[
x_{ij} = X_i + \delta_{ij} \quad \text{and} \quad y_{ik} = Y_i + \epsilon_{ik} : j = 1, 2, \ldots, m \quad \text{and} \quad k = 1, 2, \ldots, m
\]

The errors terms are distributed with Normal distribution of \( \delta_{ij} \sim N(0, \sigma^2) \) and \( \epsilon_{ik} \sim N(0, \tau^2) \). In this case, we assumed the size of elements in each group are the same, i.e. \( j = k \).

However, when a single outlier exists in the dataset, the parameter estimation will be affected. Furthermore, the observation can be considered as an outlier if it does not follow any pattern with the remainder of the other observations. In the errors-in-variable model, the issue of outliers has received
considerable critical attention that need to be investigated for example, when the observations in an experiment are incorrectly recorded or data set is mistakenly entered in the computer [8–10]. In unreplicated LFRM, the outlier detection has been established using COVRATIO statistic [11]. However, outlier detection in replicated LFRM has not been explored. Therefore, in this study, we proposed a method in detecting a single outlier using COVRATIO statistic by modifying the method that has been developed for unreplicated LFRM. Section 2 describes the maximum likelihood estimation method of replicated LFRM while in Section 3 describes the outlier detection using COVRATIO statistic. The determination of cut-off point is pointed out in Section 4. Results and Discussion in Section 5. Next, we illustrate the proposed method using a real dataset in Section 6. Lastly, the conclusion is stated in Section 7.

2. Maximum Likelihood Estimation of replicated linear functional relationship model

The estimation of parameters in replicated LFRM can be obtained by using the Maximum Likelihood Estimation (MLE) method which involves an iterative technique [2, 12]. The log-likelihood function can be shown as

\[
\log L(\alpha, \beta, \sigma^2, \tau^2, X_1, ..., X_p) = \text{constant} - \frac{1}{2} \sum_{i=1}^{p} m_i (\log \sigma^2 + \log \tau^2) - \frac{1}{2} \left\{ \sum_{i=1}^{p} \sum_{j=1}^{m_i} \frac{(x_{ij} - \bar{x}_i)^2}{\sigma^2} + \sum_{i=1}^{p} \sum_{j=1}^{m_i} \frac{(y_{ij} - \bar{y}_i - \alpha \beta x_i)^2}{\tau^2} \right\} .
\]

(3)

There are \((p + 4)\) parameters to be estimated and can be obtained by differentiating the log-likelihood function as given in equation (1) with respect to \(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2, \hat{\tau}^2\) and \(\bar{x}_i\). Thus the estimated parameters are as follows:

\[
\hat{x}_i = \frac{1}{\Delta_i} \left( \frac{m_i \bar{x}_i}{\sigma^2} + \frac{m_i \bar{y}_i}{\tau^2} (\bar{y}_i - \hat{\alpha}) \right)
\]

(4)

\[
\hat{\sigma}^2 = \frac{\sum_{i=1}^{p} \sum_{j=1}^{m_i} (x_{ij} - \hat{x}_i)^2}{\sum_{i=1}^{p} m_i} \ , \ \hat{\tau}^2 = \frac{\sum_{i=1}^{p} \sum_{j=1}^{m_i} (y_{ij} - \hat{\alpha} - \hat{\beta} x_i)^2}{\sum_{i=1}^{p} m_i}
\]

(5)

\[
\hat{\alpha} = \frac{\sum_{i=1}^{p} m_i (\bar{y}_{i} - \beta \bar{x}_i)}{\sum_{i=1}^{p} m_i} \ , \ \hat{\beta} = \frac{\sum_{i=1}^{p} m_i \bar{x}_i (\bar{y}_i - \hat{\alpha})}{\sum_{i=1}^{p} m_i \bar{x}_i^2}
\]

(6)

where \(\bar{x}_i = \frac{\sum x_{ij}}{m_i} \) and \(\bar{y}_i = \frac{\sum y_{ij}}{m_i}\) and \(\Delta_i = \frac{m_i}{\sigma^2} + \frac{m_i}{\tau^2} \).

The parameters can be solved iteratively by using unreplicated LFRM as a starting point. The estimated parameters for unreplicated LFRM are the \(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2\) and \(\hat{\tau}^2 = \hat{\sigma}^2\) where the ratio of error variances, \(\lambda = 1\) or \(\hat{\tau}^2 = \lambda \hat{\sigma}^2\) to start the iteration [11].

3. Outlier detection using COVRATIO statistic

COVRATIO statistic has been introduced in a linear regression model to identify influential observations or outliers [13]. Since then, many authors have been used COVRATIO statistic for detecting outliers because this procedure is simple, widely used and had been well established in errors-in-variable models [14–17]. By following the steps suggested in unreplicated LFRM, the COVRATIO statistic was slightly changed to accommodate replicated LFRM [11]. In replicated linear functional relationship model, the ratio of covariance statistic is suggested by

\[
|COVRATIO_{(-i)} - 1| = \left| \frac{\text{cov}_{(-i)}}{\text{cov}} \right|
\]

(7)

where \(|\text{cov}|\) is the determinant of covariance matrix for full data set and \(|\text{cov}_{(-i)}|\) is determinant of covariance matrix by the proposed method. The \(|\text{cov}_{(-i)}|\) is obtained by deleting \(i^{th}\) observation of every group and this observation is repeated with mean of every group to make balanced replication of
all sample groups. Any \(|\text{COVRATIO}_{(-i)} - 1|\) with observation exceeds the cut-off points will be considered as an outlier.

The determinant of the covariance matrix can be found using the asymptotic variances of the estimators by inverting the estimated Fisher information matrix for replicated LFRM. The covariance matrix of the parameter in the model is given by

\[
\begin{bmatrix}
Q \Sigma_{i=1}^p X_i^2 & -Q \Sigma_{i=1}^p X_i \\
-Q \Sigma_{i=1}^p X_i & Qp
\end{bmatrix}
\]

where

\[
Q = \frac{m r^2 + m \beta^2 \sigma^2}{m^2 \left( p \Sigma_{i=1}^p X_i^2 - \left( \sum_{i=1}^p y_i X_i \right)^2 \right)}.
\]

Therefore, the determinant of the covariance matrix for this model is given by

\[
|\text{COV}| = \frac{(m^2)^2}{(mr^2 + m \beta^2 \sigma^2)^2 \left( p \Sigma_{i=1}^p X_i^2 - \left( \sum_{i=1}^p y_i X_i \right)^2 \right)}.
\]

4. Determining the cut-off point

The simulation study is performed in R software to obtain the cut-off point of COVRATIO statistic for replicated linear functional relationship model. Eight different sample sizes \(n = 20, 40, 60, 80, 100, 132, 180\) and 300. Without loss of generality, the intercept, the slope and error variance parameters of replicated LFRM are fixed at \(\alpha = 1, \beta = 1\) and \(\sigma^2 = 1\) and different values of \(\tau^2 = 0.2, 0.4, 0.6, 0.8\) and 1.0 respectively. For each combination of sample size \(n\) and \(\tau^2\), the observed value of \(x_{ij}\) and \(y_{ij}\) are generated using equation (2). Subsequently, the generated data are fitted by using the parameters of replicated LFRM and the \(|\text{COV}|\) is calculated. By considering a 90% confidence level, thus the cut-off point is to be at a 10% significant level. The process is repeated 5000 times and the 10% upper percentiles of maximum value of \(|\text{COVRATIO}_{(-0)} - 1|\) is obtained. The 10% upper percentiles values are used as the cut-off points in detecting a single outlier.

5. Results and discussions

The simulation result for 10% upper percentile can be shown in table 1. From table 1, we can see that the 10% upper percentile is a decreasing function of the sample size, \(n\).

| Sample size, \(n\) | \(\tau^2 = 0.2\) | \(\tau^2 = 0.4\) | \(\tau^2 = 0.6\) | \(\tau^2 = 0.8\) | \(\tau^2 = 1.0\) |
|---------------------|----------------|----------------|----------------|----------------|----------------|
| 20                  | 0.9998         | 1.3477         | 1.7647         | 1.9923         | 2.1651         |
| 40                  | 1.0873         | 1.0473         | 1.0578         | 1.0783         | 1.1078         |
| 60                  | 0.9770         | 0.9453         | 0.9171         | 0.8924         | 0.8717         |
| 80                  | 0.9581         | 0.9084         | 0.8690         | 0.8372         | 0.8083         |
| 100                 | 0.9347         | 0.8700         | 0.8195         | 0.7797         | 0.7492         |
| 132                 | 0.8898         | 0.7987         | 0.7369         | 0.6915         | 0.6565         |
| 180                 | 0.8173         | 0.7026         | 0.6309         | 0.5855         | 0.5501         |
| 300                 | 0.6668         | 0.5330         | 0.4636         | 0.4199         | 0.3891         |
Then, the arithmetic mean of the values for each $n$ are calculated and the power series formula is plotted as shown in figure 1. From figure 1, the graph shows a good fit where the value of $R^2$ is approximately equal to 1. The equation from figure 1, $y = 5.2418 n^{-0.407}$ where $n$ is the sample size will be used as a cut-off point in detecting a single outlier. Any observations with $|\text{COVRATIO}_{(-i)} - 1|$ exceeding the cut-off point will be considered as an outlier at 10% significant level.

6. Application
The proposed method is applied to a real dataset by considering iron in a slag dataset with $n = 50$ [18]. Since the dataset represents observations taken by each of two different techniques, namely magnetic test and chemical test which are subject to measurement errors, the dataset can be considered as unreplicated data because there is only single $x$ and $y$ observation for each level of $i$ [19]. The scatter plot of the dataset can be shown in figure 2. In unreplicated LFRM, the assumption on the ratio of error variances, $\lambda$, is needed to estimate the parameters. However, in the absence of knowledge on the ratio of error variances, the data is transformed into pseudo replicates and used MLE for balanced replicated LFRM to estimate all parameters. The data is divided into 5 groups to obtain the pseudo-replicates and each group has 10 observations that are balanced and equal. Since there is no outlier in the original iron in slag data, the original data is modified as shown in figure 3 [9]. The purpose of using this dataset to test whether the proposed method can accurately detect the outlier present within the dataset. An outlier is inserted randomly namely at the $32^{nd}$ observation and then applied the proposed method. The COVRATIO statistic for each observation is calculated and any observation that exceeds the cut-off point will be considered as an outlier. When the sample size is 50, the cut-off point is $y = 5.2418 n^{-0.407} = 5.2418(50)^{-0.407} = 1.0666$ at 10% significant level. From table 2 and figure 4, the value of $32^{nd}$ observation exceeds the cut-off point and is correctly detected as an outlier of the data set through the proposed method. Hence, the developed test statistic and the cut-off points correctly detect the $32^{nd}$ observation as an outlier in the dataset.

Figure 1. Graph of the power series for the cut-off point at 10% significant level.
Figure 2. Scatter plot of original data.

Figure 3. Scatter plot of modified data.

Table 2. The values of the $|\text{COVRATIO}_{(i)} - 1|$ statistic for iron in slag data.

| Obs | $|\text{COVRATIO}_{(i)} - 1|$ | Obs | $|\text{COVRATIO}_{(i)} - 1|$ | Obs | $|\text{COVRATIO}_{(i)} - 1|$ | Obs | $|\text{COVRATIO}_{(i)} - 1|$ |
|-----|-------------------------------|-----|-------------------------------|-----|-------------------------------|-----|-------------------------------|
| 1   | 0.1648                        | 14  | 0.0399                        | 27  | 0.1625                        | 40  | 0.3956                        |
| 2   | 0.1246                        | 15  | 0.0045                        | 28  | 0.2293                        | 41  | 0.0272                        |
| 3   | 0.3812                        | 16  | 0.0726                        | 29  | 0.1041                        | 42  | 0.0585                        |
| 4   | 0.0406                        | 17  | 0.1285                        | 30  | 0.0271                        | 43  | 0.1150                        |
| 5   | 0.2314                        | 18  | 0.4900                        | 31  | 0.0344                        | 44  | 0.0611                        |
| 6   | 0.0216                        | 19  | 0.3232                        | 32  | 1.7871                        | 45  | 0.0583                        |
| 7   | 0.0065                        | 20  | 0.4501                        | 33  | 0.0982                        | 46  | 0.0670                        |
| 8   | 0.0123                        | 21  | 0.1316                        | 34  | 0.2314                        | 47  | 0.5956                        |
| 9   | 0.5669                        | 22  | 0.1525                        | 35  | 0.0647                        | 48  | 0.0812                        |
| 10  | 0.0318                        | 23  | 0.2906                        | 36  | 0.7415                        | 49  | 0.0607                        |
| 11  | 0.0067                        | 24  | 0.0096                        | 37  | 0.3753                        | 50  | 0.0078                        |
| 12  | 0.0422                        | 25  | 0.0165                        | 38  | 0.3571                        |     |                               |
| 13  | 0.0058                        | 26  | 0.1576                        | 39  | 0.0176                        |     |                               |
7. Conclusion
In conclusion, the COVRATIO statistic can be used for detecting a single outlier at 90% confidence level for replicated LFRM. The cut-off point is developed at 0.10 significant level by a simulation study. The COVRATIO statistic is used because it is a simple procedure, widely used and easy to implement to conform with replicated LFRM. As an illustration, the proposed method has been applied to a real dataset and can be used to identify a single outlier present in the data set.

8. Acknowledgement
We are most grateful to University of Malaya and National Defence University of Malaysia for supporting this work. We also wish to thank to referee for their helpful comments and suggestions.

9. References
[1] Dorff M and Gurland J 1961 J. R. Stat. Soc. Ser. B 23 (1) 160–170
[2] Mohd Arif A, Zubairi Y Z and Hussin A G 2020 Malaysian J. Fundam. Appl. Sci. 16 (2) 158–160
[3] Chan L K and Mak T K 1979 J. Multivar. Anal. 9 304–313
[4] Dolby G, Cormack R and Sinclair D 1987 Biometrika 74 (2) 393–399
[5] Hussin A G, Fieller N and Stillman E 2005 J. Appl. Sci. 5 (1) 138–143
[6] Mokhtar N A, Zubairi Y Z, Hussin A G and Yunus R M 2017 Matematika 33 (2) 159–163
[7] Alshargawi A S, Hussin A G and Abd Rauf U F 2019 Asian J. Appl. Sci. 7 (1) 10–18
[8] Abdullah M 1995 Commun. Stat. - Theory Methods 24 (6) 1585–1595
[9] Kim M G 2000 J. Appl. Stat. 27 (4) 451–460
[10] Gençay R and Gradojevic N 2011 J. Stat. Comput. Simul. 81 (11) 1545–1564
[11] Ghapor A A, Zubairi Y Z, Mamun A S M A and Imon A H M R 2014 Pak. J. Stat. 30 (1) 129–142
[12] Barnett V 1970 Appl. Stat. 19 (2) 135–144
[13] Belsley D A, Kuh E, and Welsch R E 1980 Identifying influential data and sources of collinearity (New York: John Wiley & Sons)
[14] Hussin A G, Abuzaid A, and Mohamed I 2009 Int. Conf. on Nonparametric Methods for Meas. Error Model. Relat. Top. (Canada: Ottawa) pp 1-18
[15] Hussin A G and Abuzaid A 2012 Pak. J. Stat. 28 (2) 205–216
[16] Mamun A S M A, Zubairi Y Z, Hussin A G, Imon A H M R, Rana S and Carrasco J 2019 Commun.
Stat. - Simul. Comput. 1–12
[17] Mokhtar N M, Zubairi Y Z, Hussin A G and Moslim N H 2019 Malaysian J. Sci. 38 (2) 46–54
[18] Hand D J, Daly F, Lunn A D, McConway K J, and E Ostrowski 1994 A Handbook of Small Data Sets (London: Chapman and Hall)
[19] Fah C Y, Hussin A G and Rijal O M 2007 J. Appl. Sci. 7 (1) 20–26