Sains Malaysiana 51(2)(2022): 599-607
http://doi.org/10.17576/jsm-2022-5102-23

Outlier Detection in Balanced Replicated Linear Functional Relationship Model

(Pengesanan Data Terpencil dalam Model Hubungan Fungsian Linear Bereplika Seimbang)

AZURAINI MOHD ARIF, YONG ZULINA ZUBAIRI* & ABDUL GHAPOR HUSSIN

ABSTRACT

Identification of outlier in a dataset plays an important role because their existence will affect the parameter estimation. Based on the idea of COVRATIO statistic, we modified the procedure to accommodate for replicated linear functional relationship model (LFRM) in detecting the outlier. In this replicated model, we assumed the observations are equal and balanced in each group. The derivation of covariance matrices using Fisher Information Matrices is also given for balanced replicated LFRM. Subsequently, the cut-off points and the power of performance are obtained via a simulation study. Results from the simulation studies suggested that the proposed procedure works well in detecting outliers for balanced replicated LFRM and we illustrate this with a practical application to a real data set. The implication of the study suggests that with some modification to the procedures in COVRATIO, one could apply such a method to identify outliers when modelling balanced replicated LFRM which has not been explored before.

Keywords: Covariance matrix; covratio; influential observation; linear functional relationship model; outliers

ABSTRAK

Pengesanan data terpencil di dalam set data adalah penting kerana kewujudannya akan mengganggu penganggaran nilai parameter. Berdasarkan idea statistik COVRATIO, kami mengubah suai prosedur tersebut supaya bersesuaian bagi model hubungan fungsian linear bereplika dan seimbang dalam pengesanan data terpencil. Setiap unsur dalam kumpulan adalah sama dan seimbang dalam model replikasi ini. Pembentukan matriks kovarians melalui matrik maklumat Fisher juga diberikan bagi model ini. Seterusnya, titik potongan dan kuasa prestasi bagi kaedah yang dicadangkan diperoleh melalui kajian simulasi. Hasil keputusan daripada kajian simulasi menunjukkan prosedur yang dicadangkan berfungsi dengan baik dalam pengesanan data terpencil untuk model hubungan fungsian linear bereplika dan seimbang dan kami memberikan contoh ke atas set data sebenar. Implikasi daripada kajian ini menunjukkan bahawa kita boleh mengesan data terpencil dengan sedikit pengubahsuaian terhadap prosedur COVRATIO bagi model hubungan fungsian linear bereplika dan seimbang kerana pengesanan data terpencil menggunakan model ini masih belum lagi diteroka.

Kata kunci: Covratio; data berpengaruh; matriks kovarians; model hubungan fungsian linear; terpencil

INTRODUCTION

The linear functional relationship model (LFRM) is one of the families in the error-in-variables model (EIVM) besides the linear structural relationship model and ultrastructural relationship model. Many authors have considered LFRM over the years in fitting the parameters (Barnett 1970; Cheng & Van Ness 1994; Kendall & Stuart 1979; Lindley 1947). Furthermore, the LFRM can be extended to unreplicated and replicated functional relationship models, with certain recommendation (Dorff & Gurland 1961). In unreplicated LFRM, the assumption on the ratio of error variances, $\lambda = \frac{\sigma^2}{\tau^2}$, is needed to estimate the parameters and usually equal to 1. However, in the absence of knowledge on the ratio of error variances, λ, the data can be transformed into pseudo replicates which is called as replicated and used maximum likelihood method to estimate all parameters (Hussin et al. 2005).

However, the estimation of the parameter becomes inconsistent when outliers occur in the dataset. Outliers are observations in the dataset which follow unusual patterns and occur because of gross measurement and recording errors (Aggarwal 2013). As mentioned by Hampel et al. (1986), "A routine data set typically contains about 1-10% outliers and even the highest quality data set cannot be guaranteed free of outliers". Hence, to evaluate their influence on the model, it is important to

locate the outliers. The presence of outliers changes the parameter estimates for example in the linear regression for circular variables when estimating the parameters on a wind direction dataset (Hussin et al. 2013; Rambli et al. 2015) and also on an eye dataset (Alkasadi et al. 2019).

Extensive works on outlier detection have been well established in linear models (Cheng & Van Ness 1994; Ibrahim et al. 2013; Satman et al. 2021; Wong 1989). Applications using COVRATIO statistics can be seen in physiological, epidemiology, medicine and many different disciplines (Alcaraz-Ibáñez et al. 2021; Satari & Khalif 2020; Viechtbauer & Cheung 2010). Several researchers proposed a group deleted version to identify outliers based on the COVRATIO statistic because this procedure is simple, widely used and had been well established in both linear and circular regression modelling (Belsley et al. 1980; Rambli et al. 2016). As for errors-in-variable models, outlier detection using COVRATIO statistics has been developed for the linear model (Ghapor et al. 2014; Mamun et al. 2019) as well as for circular variables (Hussin et al. 2010; Mokhtar et al. 2019). Additionally, the COVRATIO method has been used in detecting the outlier in unreplicated linear functional relationship model (Ghapor et al. 2014). However, works on identifying outliers in replicated linear functional relationship model are somewhat limited; this could largely due to the complexity of the model with little use in modelling real data sets. Nevertheless, this does not deem the model less important; in fact, this is the motivation of this study. Thus, in this article, we propose the COVRATIO statistic in detecting a single outlier in balanced replicated LFRM and investigate on the suitability of the procedure.

This paper is organized as follows: Firstly, we review the replicated linear functional relationship model and derive the covariance matrix of the balanced replicated LFRM. Secondly, we present the COVRATIO statistic in identifying the outlier. Thirdly, the procedures are described by simulation study to determine the cut-off point. Next, the performance of the statistic proposed is investigated. Finally, we illustrate the detection of the outlier using examples of some data.

REPLICATED LINEAR FUNCTIONAL RELATIONSHIP MODEL

In replicated LFRM, suppose x_{ij} and y_{ij} are the observed values of the linear variables X_i and Y_i. For any fixed X_i, we assume that may be replicated observations of X_i and Y_i occurring in p groups and measured with errors δ_{ij} and ε_{ij}, respectively. This can be written as

$$x_{ij} = X_i + \delta_{ij} \text{ and } y_{ij} = Y_i + \varepsilon_{ij}, \quad Y_i = \alpha + \beta X_i \text{ for } i = 1, \dots, p \text{ and } j = 1, \dots, m \tag{1}$$

where α is the intercept and β is the slope parameters, respectively. We assume the error terms δ_{ij} and ε_{ij} follow a normal distribution where $\delta_{ij} \sim N(0,\sigma^{\wedge 2})$ and $\varepsilon_{ij} \sim N(0,\tau^2)$, respectively (Mohd Arif et al. 2020). For balanced replicated LFRM, we assumed the elements in the groups are equal and balanced which is m. For example, if the dataset consists of 60, then the data can be divided randomly into 6 groups to obtain the pseudo-replicates and each group has 10 observations or elements that are balanced and equal. In this case, $p = 6$ and $m = 10$.

The common method that is frequently used in estimating the parameters in a replicated linear functional relationship model is the Maximum Likelihood Estimation (Barnett 1970; Hussin 2005). In the case of balanced replicated LFRM, the log-likelihood function can be expressed as

$$\log log\, L\left(\alpha, \beta, \sigma^2, \tau^2, X_1, \dots, X_p\right) = constant - \frac{1}{2}\sum_{i=1}^{p} m\left(\log log\, \sigma^2 + \log log\, \tau^2\right) - \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{m}\frac{\left(x_{ij}-X_i\right)^2}{\sigma^2} - \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{m}\frac{\left(y_{ij}-\alpha-\beta X_i\right)^2}{\tau^2} \tag{2}$$

There are $p + 4$ parameters to be estimated and can be obtained by the first partial derivative of the log-likelihood function as given in (2) with respect to $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2, \hat{\tau}^2$ and $\hat{X}_i$ respectively, and equating to zero (Barnett 1970). From Barnett (1970), by setting the equal size in each group, m, the estimates of $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2, \hat{\tau}^2$ and $\hat{X}_i$ can be solved iteratively by setting initial estimates from the unreplicated linear functional relationship model by assuming $\lambda = 1$ or $\sigma^2 = \tau^2$ or until all parameters converge. Thus, we can obtain the parameters in the order of $\hat{X}_i$ followed by $\hat{\sigma}^2$ or $\hat{\tau}^2$ and lastly $\hat{\alpha}$ or $\hat{\beta}$ as given by

$$\hat{X}_i = \frac{1}{\hat{\Delta}_i}\left\{\frac{m\bar{x}_{i.}}{\hat{\sigma}^2} + \frac{m\hat{\beta}}{\hat{\tau}^2}\left(\bar{y}_{i.} - \hat{\alpha}\right)\right\}, \tag{3}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{p}\sum_{j=1}^{m}\left(x_{ij}-\hat{X}_i\right)^2}{\sum_{i=1}^{p} m}, \quad \hat{\tau}^2 = \frac{\sum_{i=1}^{p}\sum_{j=1}^{m}\left(y_{ij}-\hat{\alpha}-\hat{\beta}\hat{X}_i\right)^2}{\sum_{i=1}^{p} m}, \tag{4}$$

$$\hat{\alpha} = \frac{\sum_{i=1}^{p} m\left(\bar{y}_{i.}-\hat{\beta}\hat{X}_i\right)}{\sum_{i=1}^{p} m}, \quad \hat{\beta} = \frac{\sum_{i=1}^{p} m\hat{X}_i\left(\bar{y}_{i.}-\hat{\alpha}\right)}{\sum_{i=1}^{p} m\hat{X}_i^2} \tag{5}$$

where $\bar{x}_{i.} = \frac{\sum x_{ij}}{m_i}$, $\bar{y}_{i.} = \frac{\sum y_{ik}}{m_i}$ and $\hat{\Delta}_i = \frac{m_i}{\hat{\sigma}^2} + \frac{m_i\hat{\beta}^2}{\hat{\tau}^2}$. Thus, for balanced replicated LFRM, all parameters can be estimated.

Next, we derived the asymptotic covariance of parameters for balanced replicated LFRM using the Fisher Information matrix. By considering the first partial derivative and minus the expected value of the second partial derivative of the log-likelihood function, we obtain the estimated Fisher information matrix, F, for $\hat{X}_1, \cdots, \hat{X}_p, \hat{\sigma}^2, \hat{\tau}^2, \hat{\alpha}$ and $\hat{\beta}$ given by

$$F = \begin{bmatrix} B & 0 & E \\ 0 & C & 0 \\ E^T & 0 & D \end{bmatrix}$$

where B is a $p \times p$ diagonal matrix having same elements equal to $\frac{m_i}{\sigma^2} + \frac{m_i\beta^2}{\tau^2}$ while E is a $p \times 2$ matrix given by $E = \begin{bmatrix} \frac{m_1\beta}{\tau^2} & \frac{m_1X_1\beta}{\tau^2} \\ \vdots & \vdots \\ \frac{m_p\beta}{\tau^2} & \frac{m_pX_p\beta}{\tau^2} \end{bmatrix}$, C is a 2×2 matrix given by $= \begin{bmatrix} \frac{n}{2\sigma^4} & 0 \\ 0 & \frac{n}{2\tau^4} \end{bmatrix}$ and D is a 2×2 matrix given by D = $\begin{bmatrix} \frac{mp}{\tau^2} & \frac{m\sum_{i=1}^{p} X_i}{\tau^2} \\ \frac{m\sum_{i=1}^{p} X_i}{\tau^2} & \frac{m\sum_{i=1}^{p} X_i^2}{\tau^2} \end{bmatrix}$, respectively.

The asymptotic covariance matrix of $\hat{\sigma}^2, \hat{\tau}^2, \hat{\alpha}$ and $\hat{\beta}$ is the bottom right minor of order 4×4 of the inverse of matrix F which is our main interest. From the theory of partitioned matrices (Graybill 1961), this is given by,

$$\widehat{Var}\begin{bmatrix} \hat{\sigma}^2 \\ \hat{\tau}^2 \\ \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} C^{-1} & 0 \\ 0 & (D - E^TB^{-1}E)^{-1} \end{bmatrix}$$

It can be shown that for $(D - E^T B^{-1} E)^{-1}$ is as follows:

$$(D - E^TB^{-1}E)^{-1} = \frac{m\tau^2 + m\beta^2\sigma^2}{m^2\left\{p\sum_{i=1}^{p} X_i^2 - \left(\sum_{i=1}^{p} X_i\right)^2\right\}}\begin{bmatrix} \sum_{i=1}^{p} X_i^2 & -\sum_{i=1}^{p} X_i \\ -\sum_{i=1}^{p} X_i & p \end{bmatrix}$$

In particular, we have the following results:

$$\hat{Var}(\hat{\alpha}) = \frac{(m\tau^2 + m\beta^2\sigma^2)\sum_{i=1}^{p} X_i^2}{m^2\left\{p\sum_{i=1}^{p} X_i^2 - \left(\sum_{i=1}^{p} X_i\right)^2\right\}} \quad (6)$$

$$\hat{Var}(\hat{\beta}) = \frac{(m\tau^2 + m\beta^2\sigma^2)p}{m^2\left\{p\sum_{i=1}^{p} X_i^2 - \left(\sum_{i=1}^{p} X_i\right)^2\right\}} \quad (7)$$

$$\hat{Cov}(\hat{\alpha}, \hat{\beta}) = -\frac{(m\tau^2 + m\beta^2\sigma^2)\sum_{i=1}^{p} X_i}{m^2\left\{p\sum_{i=1}^{p} X_i^2 - \left(\sum_{i=1}^{p} X_i\right)^2\right\}}. \quad (8)$$

COVRATIO STATISTIC FOR BALANCED REPLICATED LFRM

The COVRATIO statistic has been used in detecting the outliers in the linear regression model and also in unreplicated LFRM (Belsley et al. 1980; Ghapor et al. 2014). The idea of COVRATIO statistic is based on the determinantal ratio between determinant of the covariance matrix for a full data set and a reduced data set (Belsley et al. 1980). This is given by

$$|COVRATIO_{(-i)} - 1| = \frac{|COV|}{|COV_{(-i)}|}$$

where $|COV|$ is the determinant of the covariance matrix for a full data set and $|COV_{(-i)}|$ is the determinant for the reduced data set by excluding the i^{th} observation. If the ratio is close to one, then the i^{th} observation is consistent with other observations. In this study, the idea is extended to the replicated LFRM in which the cut-off values and formula will be derived.

For balanced replicated LFRM, the ratio of statistic is suggested by a slightly different procedure. The proposed procedure is based on the determinant of the asymptotic covariance of the parameters from (8). Although the $|COV|$ is the determinant of the covariance matrix for a full data set, the $|COV^*_{(-i)}|$ is obtained by deleting i^{th} observation of every group and this observation is repeated with the mean of every group to make balanced replication of all sample groups. The use of mean substitution may be based on the fact that the mean is a reasonable guess of a value for a randomly selected observation from a normal distribution (Acock 2005). This is given by

$$\left|COVRATIO_{(-i)} - 1\right| = \frac{|COV|}{\left|COV^*_{(-i)}\right|} \quad (9)$$

where $|COV^*_{(-i)}|$ is a determinant of the covariance matrix by the proposed method. Any $|COVRATIO_{(-i)} - 1|$ with observation exceeds the cut-off points will be considered as an outlier. As mentioned earlier, for the balanced replicated LFRM, the cut-off points are obtained through the simulation studies by following the idea from (Ghapor et al. 2014; Mamun et al. 2019).

TABLE 1. Values of sample size, group and elements

Sample size, n	Group, p	Elements, m
20	4	5
40	5	8
60	6	10
80	8	10
100	10	10
132	11	12
180	12	15
300	15	20

DETERMINATION OF CUT-OFF POINTS FOR COVRATIO STATISTIC

We carried out a simulation study to obtain the cut-off points of COVRATIO statistic for balanced replicated LFRM. Eight different sample sizes, n = 20,40,60,80,100,132,180 and 300 are used according to the division of sample size in Table 1.

Furthermore, five values of τ^2 = 0.2,0.4,0.6,0.8, and 1.0, respectively, are used. For each sample of size n and τ^2, a set of normal random errors are generated from the normal distribution with mean 0 and variance τ^2, respectively. By adopting the steps suggested by Ghapor et al. (2014), the procedure of COVRATIO statistic in Step 6 was slightly modified to accommodate for balanced replicated LFRM. The steps are listed down in detail:

Step 1 Generate a fixed $X_i = 10\left(\frac{i}{p}\right)$ of size p, with i = 1, 2, … , p where p is the number of groups. Without loss of generality, the intercept, slope and error variance parameters of balanced replicated LFRM are fixed at α = 1, β = 1 and σ^2 = 1, respectively. *Step 2* Generate two random error terms δ_{ij} and ε_{ij} from $N(0,\sigma^2)$ and $N(0,\tau^2)$, respectively. *Step 3* Calculate the observed values of x_{ij} and y_{ij} and also the value of Y_i using (1). *Step 4* Fit the generated data to balanced replicated LFRM and estimate parameters using (3), (4) and (5), respectively. *Step 5* Find the variance-covariance matrix and calculate the $|COV|$ for all data. *Step 6* Delete the i^{th} row of every group and replicate with the mean for observation in every group from the generated sample of both x_{ij} and y_{ij} where i = 1, 2, …, p and j = 1, 2, …, m. Repeat steps 4 till steps 6 to obtain $|COV_{(-i)}|$. *Step 7* Calculate ($COVRATIO_{(-i)}$ and find the value of $|COVRATIO_{(-i)}\text{-}1|$ for all i. *Step 8* Specify the maximum value of $|COVRATIO_{(-i)}\text{-}1|$.

These steps are repeated for 5000 times for each combination of sample size n and τ^2. Then, the 5% upper percentiles of the maximum values of $|COVRATIO_{(-i)}\text{-}1|$ is calculated. This upper percentile is used as the cut-off points in identifying the outliers for the balanced replicated linear functional relationship model. Table 2 shows the value of 5% upper percentiles of each value of n and τ^2, respectively. From this table, the cut-off points show a monotonic decreasing function of sample size n.

TABLE 2. The 5% upper percentile points of $|COVRATIO_{(-i)}\text{-}1|$ at τ^2 = 0.2, 0.4, 0.6, 0.8 and 1.0

Sample size, n	0.2	0.4	0.6	0.8	1.0
20	1.1321	1.9974	2.5915	2.8700	3.0111
40	1.3591	1.2633	1.3131	1.3179	1.3985
60	0.9804	0.9529	0.9281	0.9078	0.8913
80	0.9625	0.9158	0.8807	0.8498	0.8227
100	0.9401	0.8797	0.8322	0.7931	0.7632
132	0.8967	0.8086	0.7491	0.7049	0.6705
180	0.8245	0.7139	0.6431	0.5977	0.5613
300	0.6741	0.5412	0.4719	0.4288	0.3969

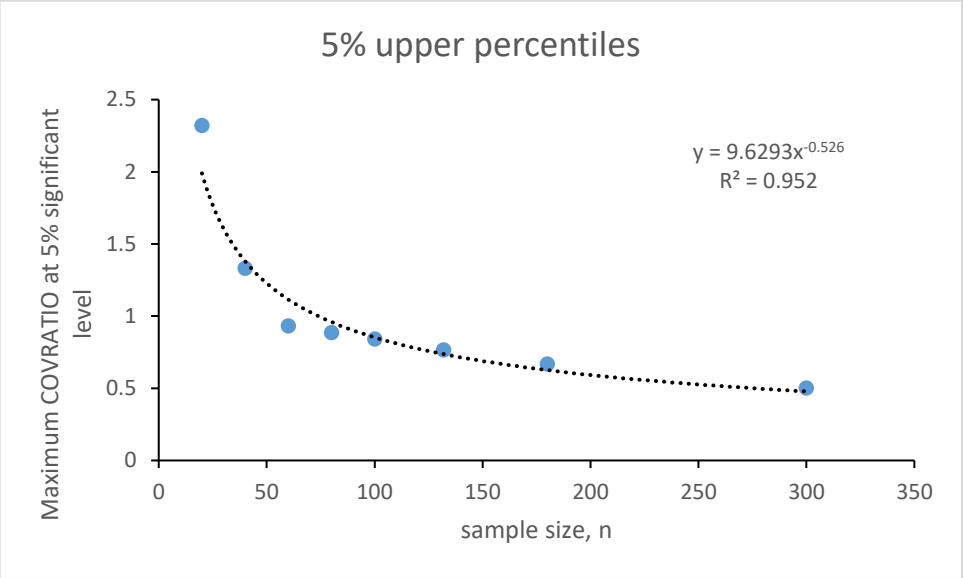

FIGURE 1. Graph of the power series in finding the general formula for the cut-off point at 5% significant level

The arithmetic mean of the values for the respective n are calculated and the best fit is obtained by using the least squares method as in Figure 1. For 5% significant level, we obtain the equation of the series trend line line $y = 9.6293n^{-0.526}$. This trend line will be used as a cut-off point in detecting the outliers. The next step of the study is to determine the power performance of the proposed method.

POWER OF PERFORMANCE FOR COVRATIO STATISTIC

Again, a Monte Carlo simulation method is used to investigate the performance of $|COVRATIO_{(-i)}\text{-}1|$ in identifying the outlier in the balanced replicated LFRM. Four different sample sizes of 40, 80, 100, and 180 are considered in this study by using the procedures described earlier to generate the data set. The power of performance $|COVRATIO_{(-i)}\text{-}1|$ are tested when correctly detecting the outlier. To determine the performance of the COVRATIO statistic, contamination is randomly applied; for example at observation c, where y_c is contaminated as follows:

$$y_c = \alpha + \beta X_C + \varphi_C$$

where y_c and X_C are the value of the c^{th} observation of both variables y and X, respectively, after contamination and φ_C is error taken from a normal distribution with mean zero and different variances of 6,8,10,12,14, and 16, respectively (Ghapor et al. 2014; Mamun et al. 2019). The generated data are refitted and the maximum of $|COVRATIO_{(-i)}\text{-}1|$ statistic is specified. This procedure has correctly identified the outlier in the data set if the values of $|COVRATIO_{(-i)}\text{-}1|$ is maximum and exceed the stated cut-off point. The process is repeated 5000 times and the power of performance is then examined by calculating the percentage of the correct detection of the contaminated observation at cth position. Figure 2 shows the power of performance of $|COVRATIO_{(-i)}\text{-}1|$ statistics for $n = 80$ (8×10) for $\tau^2 = 0.2,0.4,0.6,0.8$, and 1.0. From this plot, it can be concluded that as τ^2 decreases, the power of performance in detecting the correct outlier increases.

Figure 3 shows the power of performance of $||COVRATIO_{(-i)}\text{-}1|$ statistics for $\tau^2 = 0.2$. We can say that the power of performance is independent of the sample size by looking at both plots as shown in Figures 2 and 3.

APPLICATION

We consider two data sets where both have 10 groups and three elements in each group. The first data set is from simulated data generated from replicated LFRM by setting the parameters $\alpha = 0, \beta = 1, \sigma^2 = 1$ and $\tau^2 = 0.2$. Next, we contaminated the observation randomly. The COVRATIO statistic for each value is calculated and the

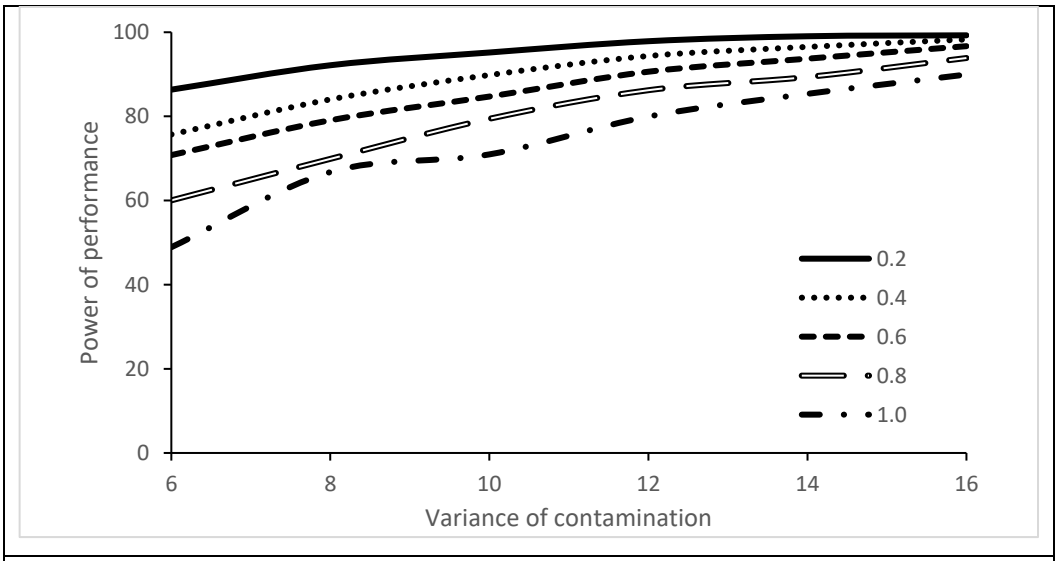

FIGURE 2. Power of performance for $|COVRATIO_{(-i)}\text{-}1|$ for $n = 80$

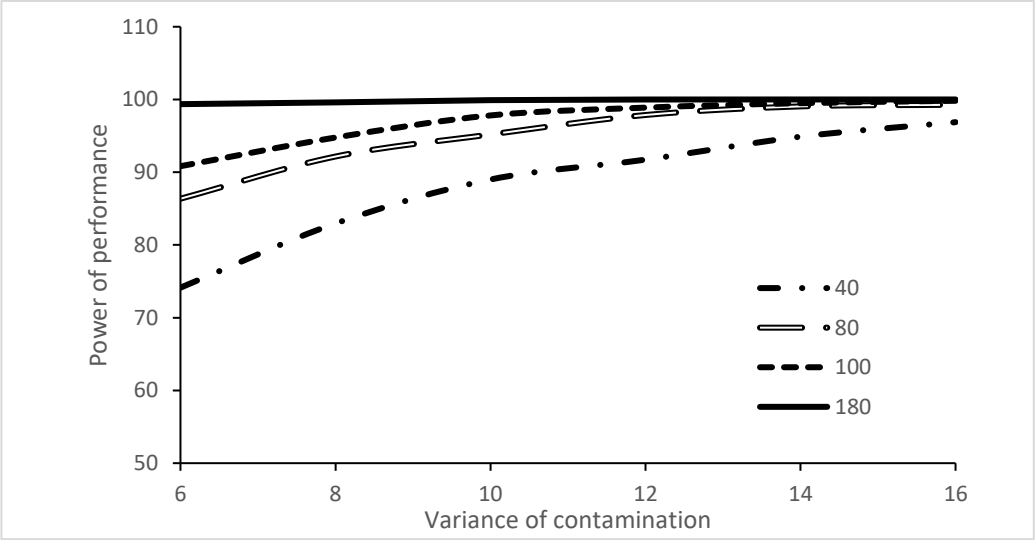

FIGURE 3. Power of performance for $|COVRATIO_{(-i)}\text{-}1|$ when $\tau^2 = 0.2$

TABLE 3. The COVRATIO statistic for simulated data

Index	$COVRATIO_{(-i)}\text{-}1\|$	Index	$COVRATIO_{(-i)}\text{-}1\|$	Index	$COVRATIO_{(-i)}\text{-}1\|$
1	0.9726	11	0.3911	21	0.4404
2	0.9760	12	0.4826	22	0.7032
3	0.9781	13	0.0230	23	0.6234
4	0.8263	14	0.2483	24	0.6850
5	0.1331	15	0.2449	25	0.9171
6	0.7490	16	0.0253	26	0.9071
7	0.3807	17	0.0068	27	0.9234
8	**7.2855**	18	0.0121	28	0.9686
9	0.4360	19	0.4510	29	0.9709
10	0.3444	20	0.2116	30	0.9635

results are given in Table 3. Based on the formulation as given in Table 2, the cut-off point for is calculated and the value 1.609 obtained as the cut-off point at 5% significant level. Figure 4 clearly shows that the COVRATIO value for 8th observation is 7.286 which exceeds the cut-off points of 1.609. Hence, the developed test statistic and the cut-off points correctly detect the 8th observation as an outlier in the simulated data set.

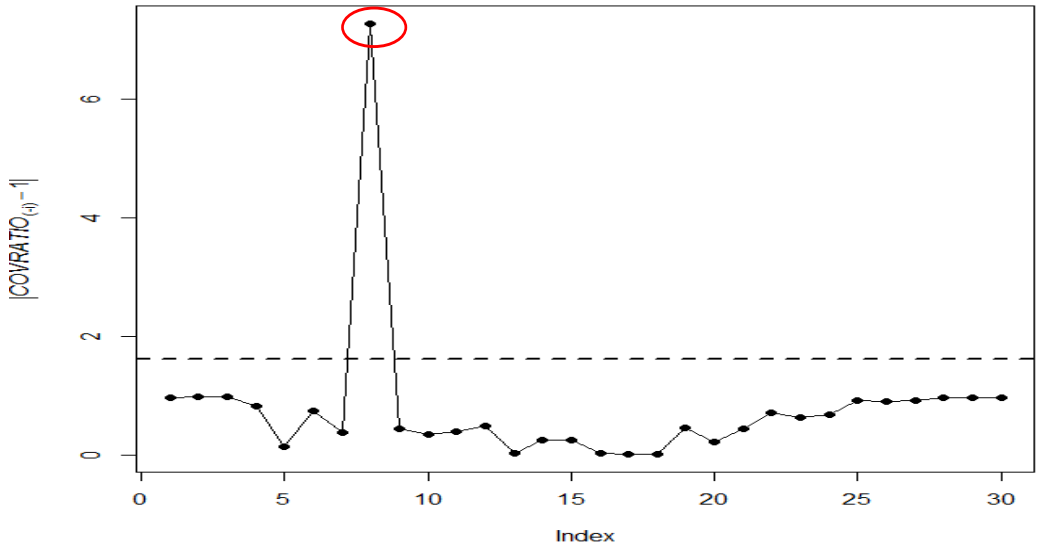

FIGURE 4. The graph of the $|COVRATIO_{(-i)}\text{-}1|$ statistic for simulated data

Next, we consider another dataset taken from Altman and Bland (1999). We use a subsample of the original data containing 30 observations. The data set measures the systolic blood pressure which simultaneous measurements were made by two experienced observers denoted as J and R. In this case, we have 10 groups (or subjects) and each group have three sets of readings that were made in quick succession. Since there is no outlier in the original data, we insert the outlier randomly into the original data by following Kim (2000) and Imon and Hadi (2008). The COVRATIO statistic for each value is calculated and the results are given in Table 4. The cut-off point for n = 30 is 1.609 obtained as before at 5% significant level. From Figure 5, we observe that the value of $|COVRATIO_{(-i)}\text{-}1|$ for the 11th observation is 1.849, which exceeds the cut-off point value of 1.609 at 5% significant level. To conclude, our cut-off point correctly identifies that the 11th observation as an outlier in this data set.

TABLE 4. The COVRATIO statistic for real data

Index	$\|COVRATIO_{(-i)}\text{-}1\|$	Index	$\|COVRATIO_{(-i)}\text{-}1\|$	Index	$\|COVRATIO_{(-i)}\text{-}1\|$
1	0.2170	**11**	**1.8490**	21	0.5978
2	0.1225	12	0.1424	22	0.0106
3	0.0987	13	1.0353	23	0.0406
4	0.0029	14	0.3322	24	0.0604
5	0.0246	15	0.3566	25	0.5508
6	0.0130	16	0.8987	26	0.0136
7	0.9255	17	0.4328	27	0.3087
8	0.9635	18	0.9015	28	0.0358
9	0.9572	19	0.8729	29	0.3755
10	0.0083	20	0.0130	30	0.1895

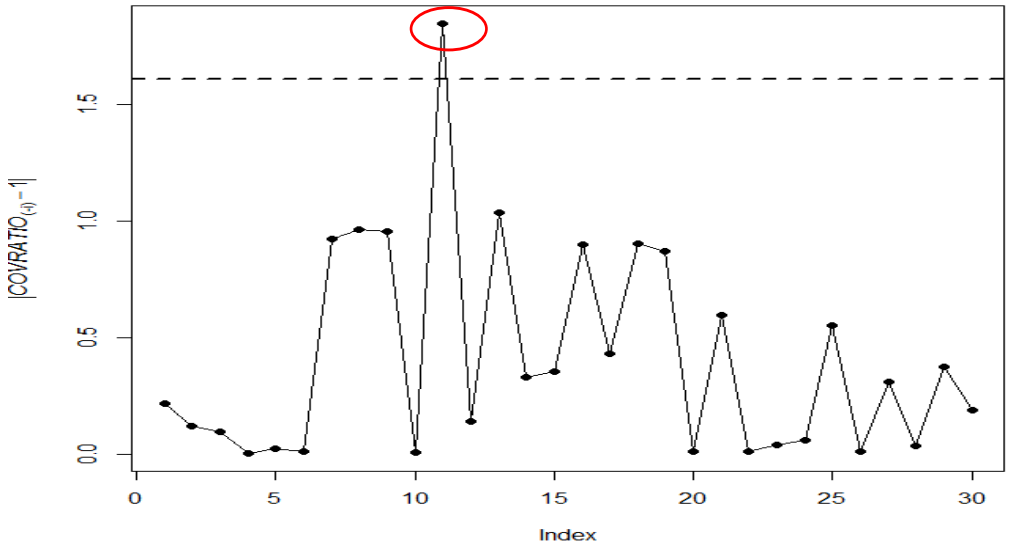

FIGURE 5. The values of the $|COVRATIO_{(-i)}-1|$ statistic for real data

CONCLUSION

We extend the idea of COVRATIO statistics for detecting a single outlier because it is a new topic and has not been explored in balanced replicated LFRM. The COVRATIO statistic is used because it is a simple procedure, widely used and easy to implement to conform with balanced replicated LFRM. In adopting the COVRATIO procedure to balanced replicated LFRM, we need to determine the cut-off point and measure the power of performance as well as a slight modification to COVRATIO to accommodate the replicates of the model. Using simulation studies, we identified the cut-off point for this statistic and have shown that this statistic performs well in identifying an outlier. A practical example is illustrated with real data set. As an implication, this study provides empirical evidence that COVRATIO can be utilized for detecting outliers on balanced replicated LFRM; thus providing a better understanding to the approach in outlier detection in the model.

ACKNOWLEDGEMENTS

We are most grateful to University of Malaya (BKS005-2019 and GPF006H-2018), National Defence University of Malaysia and Ministry of Higher Education (MoHE), Malaysia for the financial support. We also wish to thank to referees for their helpful comments and suggestions.

REFERENCES

Acock, A.C. 2005. Working with missing values. *Journal of Marriage and Family* 67(4): 1012-1028.

Aggarwal, C.C. 2013. *Outlier Analysis*. Springer: New York.

Alcaraz-Ibáñez, M., Paterna, A., Sicilia, A. & Griffiths, M.D. 2021. A systematic review and meta-analysis on the relationship between body dissatisfaction and morbid exercise behaviour. *International Journal of Environmental Research and Public Health* 18(2): 585.

Alkasadi, N.A., Ibrahim, S., Abuzaid, A.H.M., Yusoff, M.I., Hamid, H., Zhe, L.W. & Razak, A.A. 2019. Outlier detection in multiple circular regression model using DFFITc statistic. *Sains Malaysiana* 48(7): 1557-1563.

Barnett, V.D. 1970. Fitting straight lines-the linear functional relationship with replicated observations. *Applied Statistics* 19(2): 135-144.

Belsley, D.A., Kuh, E. & Welsch, R.E. 1980. Regression Diagnostics: *Identifying Influential Data and Sources of Collinearity*. United States: John Wiley & Sons. p. 300.

Cheng, C.L. & Van Ness, J.W. 1994. On estimating linear relationship when both variables are subject to errors. *Journal of Royal Statistical Society: Series B (Methodological)* 56(1): 167-183.

Dorff, M. & Gurland, J. 1961. Estimation of the parameters of a linear functional relation. *Journal of the Royal Statistical Society Series B (Methodological)* 23(1): 160-170.

Ghapor, A.A., Zubairi, Y.Z., Mamun, A.S.M.A. & Imon, A.H.M.R. 2014. On detecting outlier in simple linear functional relationship model using COVRATIO statistic. *Pakistan Journal of Statistics* 30(1): 129-142.

Graybill, F.A. 1976. *Theory and Application of the Linear Model*. North Scituate: Duxbury Press.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A. 1986. Robust Statistics: *The Approach Based on Influence Functions*. New York: John Wiley & Sons. p. 502.

Hussin, A.G. 2005. Approximating fisher's information for the replicated linear circular functional relationship model.

Bulletin of the Malaysian Mathematical Sciences Society 28(2): 131-139.

Hussin, A.G., Abuzaid, A.H., Ibrahim, A.I.N. & Rambli, A. 2013. Detection of outliers in the complex linear regression model. *Sains Malaysiana* 42(6): 869-874.

Hussin, A.A., Abuzaid, A.H., Zulkifili, F. & Mohamed, I. 2010. Asymptotic covariance and outlier detection in a liner functional relationship model for circular data with an application to the measurements of wind directions. *ScienceAsia* 36(3): 249-253.

Hussin, A.G., Fieller, N. & Stillman, E. 2005. Pseudo-Replicates in the linear circular functional relationship model. *Journal of Applied Sciences* 5(1): 138-143.

Ibrahim, S., Rambli, A., Hussin, A.A. & Mohamed, I. 2013. Outlier detection in a circular regression model using COVRATIO statistic. *Communication in Statistics - Simulation and Computation* 42(10): 2272-2280.

Imon, A.H.M.R. & Hadi, A.S. 2008. Identification of multiple outliers in logistic regression. *Communications in Statistics-Theory and Methods* 37(11): 1697-1709.

Kendall, M.G. & Stuart, A. 1979. *The Advanced Theory of Statistics*. London: Griffin. p. 684.

Kim, M.G. 2000. Outliers and influential observations in the structural errors-in-variables model. *Journal of Applied Statistics* 27(4): 451-460.

Lindley, D.V. 1947. Regression lines and the linear functional relationship. *Supplement to the Journal of the Royal Statistical Society* 9(2): 218-244.

Mamun, AA.S.M.A., Zubairi, Y.Z., Hussin, A.G., Imon, A.H.M.R., Rana, R. & Carrasco, J. 2019. Identification of influential observation in linear structural relationship model with known slope. *Communications in Statistics - Simulation and Computation:* DOI.10.1080/03610918.2019.1645172.

Mohd Arif, A., Zubairi, Y.Z. & Hussin, A.G. 2020. Parameter estimation in replicated linear functional relationship model in the presence of outliers. *Malaysian Journal of Fundamental and Applied Sciences* 16(2): 158-160.

Mokhtar, N.A., Zubairi, Y.Z., Hussin, A.G. & Moslim, N.H. 2019. An outlier detection method for circular linear functional relationship model using covratio statistics. *Malaysian Journal of Science* 38(2): 46-54.

Rambli, A., Abuzaid, A.H.M., Mohamed, I. & Hussin, A.G. 2016. Procedure for detecting outliers in a circular regression model. *PLoS ONE* 11(4): 0153074.

Rambli, A., Yunus, R.M., Mohamed, I. & Hussin, A.G. 2015. Outlier detection in a circular regression model. *Sains Malaysiana* 44(7): 1027-1032.

Satari, S.Z. & Ku Khalif, K.M. 2020. Review on outliers identification methods for univariate circular biological data. *Advances in Science, Technology and Engineering Systems* 5(2): 95-103.

Satman, M.H., Adiga, S., Angeris, G. & Akadal, E. 2021. LinRegOutliers: A Julia package for detecting outliers in linear regression. *The Journal of Open Source Software* 6(57): 1-6.

Viechtbauer, W. & Cheung, M.W.L. 2010. Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods* 1(2): 112-125.

Wong, M.Y. 1989. Likelihood estimation of a simple linear regression model when both variables have error. *Biometrika* 76(1): 141-148.

Azuraini Mohd Arif
Institute of Advanced Studies
University of Malaya
50603 Kuala Lumpur, Federal Territory
Malaysia

Azuraini Mohd Arif
Centre for Foundation Studies
National Defence University of Malaysia
57000 Kuala Lumpur, Federal Territory
Malaysia

Yong Zulina Zubairi*
Centre for Foundation Studies in Science
University of Malaya
50603 Kuala Lumpur, Federal Territory
Malaysia

Abdul Ghapor Hussin
Faculty of Defence Science and Technology
National Defence University of Malaysia
57000 Kuala Lumpur, Federal Territory
Malaysia

*Corresponding author; email: yzulina@um.edu.my

Received: 2 February 2021
Accepted: 19 June 2021