MJFAS
Malaysian Journal of Fundamental and Applied Sciences

Ismaeel *et al*. | Malaysian Journal of Fundamental and Applied Sciences, Vol. 17 (2021) 636-646

RESEARCH ARTICLE

Robust Multicollinearity Diagnostic Measure for Fixed Effect Panel Data Model

Shelan Saied Ismaeel[a,*], Habshah Midi[b], Muhammed Sani[c]

[a] Department of Mathematics, Faculty of Science, University of Zakho, Iraq;
[b] Department of Mathematics and Statistics, Faculty of Science and Institute for Mathematical Research, Universiti Putra Malaysia, Serdang, Selangor Malaysia;
[c] Department of Mathematical Sciences, Federal university Dustin-MA, Katsina State, Nigeria

Abstract It is now evident that high leverage points (HLPs) can induce the multicollinearity pattern of a data in fixed effect panel data model. Those observations that are responsible for this phenomenon are called high leverage collinearity-enhancing observations (HLCEO). The commonly used within group ordinary least squares (WOLS) estimator for estimating the parameters of fixed effect panel data model is easily affected by HLCEOs. In their presence, the WOLS estimates may produce large variances and this would lead to erroneous interpretation. Therefore, it is imperative to detect the multicollinearity which is caused by HLCEOs. The classical Variance Inflation Factor (CVIF) is the commonly used diagnostic method for detecting multicollinearity in panel data. However, it is not correctly diagnosed multicollinearity in the presence of HLCEOs. Hence, in this paper three new robust diagnostic methods of diagnosing multicollinearity in panel data are proposed, namely the RVIF (WGM-FIMGT), RVIF (WGM-DRGP) and RVIF (WMM) and compared their performances with the CVIF. The numerical evidences show that the CVIF incorrectly diagnosed multicollinearity but our proposed methods correctly diagnosed no multicollinearity in the presence of HLCEOs where RVIF (WGM-FIMGT) being the best method as it has the least computational running time.

Keywords: High Leverage Points, Generalized-M estimator, High Leverage Collinearity Enhancing Observations, Multicollinearity, Within Group Least Squares Estimator.

1. Introduction

Multicollinearity describes a near-linear dependency of two or more predictor variables. In the presence of multicollinearity, the commonly used ordinary least squares (OLS) method may have wrong sign problem and produces estimates with large variances and this would lead to inaccurate prediction. Multicollinearity occurs as a result of data collection method employed, constraints on the model, model specification and overdetermined model.

It is now evident that high leverage points which are outliers in the X-space may be another source of multicollinearity (Kamruzzaman and Imon,2002; Bagheri and Midi ,2011; Bagheri et al. ,2012; Midi et al. ,2011). This recent source of multicollinearity was considered as a new case of collinearity-influential observations. According to Bagheri et al.(2012) , high leverage collinearity-influential observations (HCIO) are those observations that can changed the pattern of multicollinearity in a data set (Midi et al., 2018; Bagheri et al., 2011; Habshah et al., 2009; Hadi, 2011; Sengupta and Behimasankaram, 1997). HCIO can be classified into high leverage collinearity-enhancing (HLCEO) and high leverage collinearity-reducing (HLCRO)

***For correspondence:** shelan.ismaeel@uoz.edu.krd

Received: 19 August 2021
Accepted: 14 October 2021

observations. Those high leverage points that induce or reduce multicollinearity in a data set is called HLCEO and HLCRO, respectively. HLCEO refers to HLPs which caused multicollinearity in their presence. The OLS estimates will be much affected by their presence. Hence it is crucial to detect multicollinearity which is caused by HLPs. Classical Variance Inflation Factor (CVIF) is the commonly used diagnostic method to diagnose multicollinearity in multiple linear regression model. However, Bagheri and Midi (2011) highlighted that CVIF is not very successful in detecting multicollinearity in the presence of HLPs. They introduced robust VIFs which are based on MM and Diagnostic Robust Generalized Potential (DRGP) of Habshah et. al. (2009) and showed that both methods are successful in diagnosing multicollinearity in multiple linear regression model.

Unfortunately, to the best of our knowledge, we have not encountered any article in the literature that discusses diagnostic methods for diagnosing multicollinearity in fixed effect panel data model. Panel data is a multi-dimensional data in which many individuals are measure over time. It contains multiple observations over multiple time periods for the same firms or individuals. Since panel data model also involves more than one independent variables, it is very likely that these variables are correlated with each other resulting to a multicollinearity problem. Moreover, HLCEO may also cause multicollinearity in panel data model. Like linear model, as anticipated the presence of multicollinearity in panel data set will affect the efficiency of the parameter estimation. Therefore, it is very crucial to detect multicollinearity in panel data, specifically in the presence of HLCEO so that suitable actions can be taken to remedy such problems. This paper is organized as follows; the classical within group estimator based on the ordinary least squares method for fixed effect panel data model is presented in Section 2. The Robust Within Group Estimator based on GM-FIMGT for Fixed Effect Panel Data Model is presented in Section 3. The new proposed robust VIF in panel data model is presented in Section 4. Simulation study and numerical examples are discussed in Section 5 and 6, respectively. Finally, the concluding remark is given in Section 7.

2. Classical within group estimator based on the ordinary least squares method for fixed effect panel data model.

Consider the fixed effect panel data model (see Bramati and Croux, 2007) ;

$$y_{it} = \alpha_i + x_{it}'\beta + \varepsilon_{it} \ , \tag{1}$$

for $i = 1,2,\ldots,n$ individual units observed at time *t = 1,2,…T* . y_{it} is the response variable, α_i are the unobservable time-invariant individual effects and considered to be fixed, β is $k \times 1$ vector of regression parameter and x_{it} is *k x 1* column vector of *k*-explanatory variables. The ε_{it} denotes the error terms which are assumed to be uncorrelated across time and individual units. Within each time series, panel data needs to be firstly transformed before any statistical method is applied to the transformed data. The commonly used method to transform the data is by using mean centering (see Baltagi, 2008) as follows:

$$\begin{aligned} \tilde{y}_{it} &= y_{it} - \tfrac{1}{T}\textstyle\sum_{t=1}^{T} y_{it} \\ \tilde{x}_{it}^{(k)} &= x_{it}^{(k)} - \tfrac{1}{T}\textstyle\sum_{t=1}^{T} x_{it}^{(k)} \end{aligned} \tag{2}$$

Greene (2007) highlighted that by performing this centering operation will eliminate any unobserved time-invariant individual fixed effects α_i and Equation (1) becomes

$$\tilde{y}_{it} = \beta' \tilde{x}_{it} + \varepsilon_{it} \tag{3}$$

The classical Within Group estimator $\hat{\beta}_{WG}$, denoted as WOLS is obtained when regressing $\tilde{y}_{it}$ on $\tilde{x}_{it}$ by Ordinary Least Squares (OLS) method.

3. Robust within group estimator based on GM-FIMGT for fixed effect panel data model.

Since the WOLS easily affected by outliers, Bramati and Croux (2007) introduced Robust Within Group GM estimator based on GM6 (WGM-GM6) estimator using median centering. According to Bakar and Midi [2015a,2015b], using median centering produces nonlinearity to the resulting data and make the equivariance properties of the robust estimator redundant. Moreover, Bagheri and Midi (2015a) highlighted the weakness of GM6 is that it suffers from swamping effect. Furthermore, Ismaeel and Midi (2018) noted that GM6 has long computational running times due to using Robust Mahalanobis distance (RMD) which is based on minimum volume ellipsoid (MVE). Ismaeel and Midi (2018) pointed out that the efficiency of the GM6 estimates tend to decrease as the number of good leverage points increases. Hence, they developed a fast GM estimator which is based on Index Set Equality (ISE) which is much faster than MVE as noted by Lim and Midi (2016). We refer to this newly developed GM estimator as GM-FIMGT. The Within group GM-FIMGT estimator denoted as WGM-FIMGT is then developed by applying the GM-FIMGT to the transformed data whereby transformation of data within each time series is done by using MM-centering which is more robust than the median centering. The main attraction of WGM-FIMGT is that it only down weight vertical outliers and bad leverage points unlike the GM6 estimator where it down weights all detected high leverage points irrespective of whether they are good or bad leverage points. As noted by Rousseeuw and Van Zomeren (1990), good leverage points may contribute to the efficiency of parameter estimates and should not be down weighted. This is the main weakness of the WGM-GM6 estimator.

The GM estimator is defined as the solution of normal equations (see Maronna et al., 2006, Krasker and Welsch, 1982) which is given by:

$$\sum_{i=1}^{n} \pi_i \, \psi\left(\frac{y_i - x_i^t \hat{\beta}}{\hat{\sigma}\pi_i}\right) x_i = 0 \tag{4}$$

where $\psi = \rho'$ is a derivative of redescending function (weight function) and $\pi_i, i = 1,2, \dots, n$ is the ith initial weight element of the diagonal matrix W, $\hat{\sigma}$ is the scale estimate, and $\hat{\beta}$ is the vector of parameters estimates. The main aim of GM estimator is to down weight high leverage points which have large residuals.

It is important to highlight that most GM estimators depend on the high leverage points diagnostic methods of detection of HLPs. The GM6 utilizes the Robust Mahalanobis Distance (RMD) based on Minimum Volume Ellipsoid (MVE) or Minimum Covariance Determinant (MCD) for the identification of high leverage points.

The i^{th} vector of explanatory variables can be written as:

$x_i' = (1, x_1, x_2, \dots, x_P) = (1, t_i)$,

where t_i is a p –dimensional row vector. We calculate the mean vector $\bar{t} = {}^1\!/_n \sum_{i=1}^{n} t_i$ and the variance- covariance matrix

$C = \left(\frac{1}{n-1}\right) \sum_{i=1}^{n} (t_i - \bar{t}\,)'(t_i - \bar{t})$.

Then, Mahalanobis distance (MD) for each point is defined as follows:

$$MD_i = \sqrt{(t_i - T(X)')C(X)^{-1}(t_i - T(X)')^t} \qquad i = 1,2, \dots, n, \tag{5}$$

where $T(X)$ is the mean vector ($\bar{t}$) and $C(X)$ is the variance covariance matrix (C).

Since the location and scatter estimates in (5) are not robust, Rousseeuw and Leroy (1987) proposed utilizing Robust Mahalanobis Distance (RMD) as a diagnostic tool for identifying HLPs by substituting the classical mean vector, $T(X)$ and classical covariance matrix, $C(X)$ of MD_i in Equation [5] by robust estimators based on Minimum Volume Ellipsoid (MVE) or Minimum Covariance Determinant (MCD).

The RMD (see Habshah et al., 2009; Rousseeuw and Leroy, 1987) is then defined as follows:

$$RMD_i = \sqrt{(t_i - T(X)')V_R^{-1}(t_i - T(X)')^t},\ i = 1,2,\dots,n$$

where $\overline{\mathrm{X}}_{\mathrm{R}}$ and V_{R} are robust locations and scatter estimates based on the MVE, respectively.

The choice of initial weight π_i is very important. The initial weight of GM6 estimator (Coakley and Hettmansperger , 1993) is formulated based on RMD which is given by:

$$\pi_i = \min\left[1, \left(\frac{\chi^2_{(0.95,p)}}{RMD^2}\right)\right], i = 1,2,\dots,n \tag{6}$$

We have noted earlier, the main weakness of GM6 where it utilizes RMD based on MVE which has swamping effect, long computational running time and its' efficiency decreases as the number of good leverage points increases. Hence, this initial weight is not very efficient in the presence of good leverage points.

The shortcoming of GM6 estimator has inspired Ismaeel and Midi (2018) to propose a new GM estimator which is based on Fast Improvised Generalized Studentized Residual (FIMGT) which utilizes RMD based on Index Set Equality (ISE). We call this method fast because the ISE needs much less times for computing the mean and covariance matrix as shown by Lim and Midi (2016). The FIMGT able to classify observations into four portions, namely clean observations, vertical outliers, good leverage points (GLPs) and bad leverage points (BLPs). The GM-FIMGT is then developed by formulating an initial weight, π_i which guarantee only vertical outliers and bad leverage points be down weighted to increase the efficiency of the estimates.

$$\pi_i = \min\left[1, \left(\frac{\mathrm{CP}_{\mathrm{FIMGT}}}{FIMGT}\right)\right], i = 1,2,\dots,n \tag{7}$$

where the cut-off points for FIMGT denoted as $\mathrm{CP}_{\mathrm{FIMGT}}$ is defined as follows;

$$CP_{\mathrm{FIMGT}} = \mathrm{Median}(\mathrm{FIMGT_i}) + \mathrm{c\ MAD}(\mathrm{FIMGT_i}) \tag{8}$$

where c is equals to 2 or 3. The median absolute deviation (MAD) is defined (see Ismaeel and Midi, 2018; Midi et al., 2021) as $\mathrm{MAD}(\mathrm{FIMGT_i}) = \mathrm{median}\{|\mathrm{FIMGT_i} - \mathrm{median}(\mathrm{FIMGT_i})|\}/0.6745$. The initial weight π_i of the regular observation and GLPs are given weight equals 1 and the vertical outliers and bad leverage points are given weight equals $\frac{\mathrm{CP}_{\mathrm{FIMGT}}}{FIMGT}$. Due to space limitation, the GM-FIGMT estimator (for detail, see Ismaeel and Midi ,2018) is summarized as follows:

Step 1: Use the S estimator (Rousseuw, 1984) as an initial estimator to achieve a high breakdown of 50% with a $n^{-1/2}$ rate of convergence, and calculate the residuals (r_i).
Step 2: Based on the residuals in Step 1, compute the estimated scale of the residuals, denoted as s and it is defined $as\ s = (1.4826)(the\ median\ of\ the\ largest\ (n-p) of\ the\ |r_i|)$, 1.4826 is used to achieve consistency at the normal distribution.
Step 3: Using the estimated residuals (r_i) and the estimated scale (s), compute the standardized residuals (e_i), where, $e_i = r_i/s$
Step 4: Calculate the initial weight based on FIMGT, where $\pi_i = min\ [1, \frac{\mathrm{CP}_{\mathrm{FIMGT}}}{FIMGT}]$.
Step 5: Use the initial weight (step 4) and the standardized residuals (step 3) to achieve a bounded influence function for bad leverage points, $t_i = e_i/\pi_i$.
Step 6: Use the weighted residuals (t_i) in the first iteration of the weighted least squares (WLS) to estimate the parameters of the regression, $\hat{\beta} = (X^T W\ X)^{-1} X^T W\ Y$, where the weight w_i is reduced for large residuals to get good efficiency (Tukey weight function is used in this paper).
Step 7: Calculate the new residuals (r_i) from WLS and repeat steps (2-6) until convergence.
The robust Within Group estimator based on FIMGT (WGM-FIMGT) is then established by applying the GM-FIMGT to the transformed data based on MM-centering.

4. New proposed robust VIF in panel data model.

Bagheri and Midi (2011) pointed out that the CVIF which is used to diagnose multicollinearity in multiple linear regression fails to correctly detect multicollinearity in the presence of HLPs. To remedy this problem, they introduced robust VIFs which are based on MM and Diagnostic Robust Generalized Potential (DRGP) of Habshah et al.(2009). However, we have not encountered any article in the literature that discusses robust diagnostic methods for diagnosing multicollinearity in fixed effect panel data model. Since within group estimator is established by employing statistical method such as OLS on the transformed model as in Equation 3 (which is equivalent to multiple linear regression model), we would like to extend the idea of formulating robust VIF for fixed effect panel data model by using similar manner as in linear model.

The CVIF is the most popular method to identify multicollinearity and it is given by

$$VIF_j = \frac{1}{1-R_j^2}, j = 1,2,\dots,p \tag{9}$$

where R_j^2 is the coefficient of multiple determination when x_j is regressed on other $X_{(p-1)}$ variables in the model using Ordinary Least Squares (OLS) method.
Following this idea, we formulate robust VIF based on our newly developed WGM-FIMGT estimator as follows:

Step 1: Compute the location MM estimate for each dependent and independent variable of panel data.

Step 2: Transform data by using MM-centering instead of mean-centering (see Ismaeel and Midi, 2018) as:

$$\tilde{y}_{it} = y_{it} - \hat{\mu}_{mm}\{y_{it}\} \tag{10}$$
$$\tilde{x}_{it}^{(j)} = x_{it}^{(j)} - \hat{\mu}_{mm}\{x_{it}^{(j)}\}$$

for $1 \le i \le n, 1 \le t \le T$ and $1 \le j \le p$, where $x^{(j)}$ is the j-th explanatory variable.

Step 3: Regress $\tilde{y}_{it}$ on $\tilde{x}_{it}$ using Fast Improvised Generalized Studentized Residuals based on Index Set Equality (FIMGT) and obtained the parameter estimates of the fixed effect panel data model.

Step 4: Calculate $RR^2{}_{(W\text{GM-FIMGT})}$, as follows.

$$RR^2{}_{(W\text{GM-FIMGT})} = 1 - \frac{\sum_{i=1}^{n} w_i r_i^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2} \tag{11}$$

where w_i and r_i are the robust weights and residuals obtain from WGM-FIMGT, respectively. The $\bar{y}$ is the weighted average of y, calculated as

$$\bar{y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

Step 5: The RVIF based on WGM-FIMGT is formulated as below:

$$RVIF_j(WGM - FIMGT) = \frac{1}{1-RR_j^2(WGM-FIMGT)}; j = 1,2,\dots,p \tag{12}$$

where $RR_j^2(WGM - FIMGT)$ is the coefficient of multiple determination when x_j is regressed on other $X_{(p-1)}$ variables in the model using WGM-FIMGT estimator. Using similar manner, the RVIF (WMM) and RVIF (WGM-DRGP) can also be formulated. This two RVIF measures can also be considered as new robust diagnostic measure of multicollinearity in panel data model since they have not been formulated before.

5. Monte Carlo simulation study

In this section, a simulation study is done to evaluate the performance of our newly developed robust

multicollinearity diagnostic measures, RVIF(WGM-FIMGT), RVIF(WMM) and RVIF (WGM-DRGP) and to compare its performance with CVIF. Following the idea of Bakar and Midi (2015), the explanatory variables were generated as follows:

$$y_{it} = \alpha_i + x'_{it}\beta + \varepsilon_{it} \ , \tag{13}$$

where ε is the error term distributed as $\mathrm{N}(0,1)$ and the time-invariant variables were distributed as $\alpha \sim \mathrm{U}(0,1)$ while the explanatory variables are generated from a multivariate standard normal distribution $\mathrm{N}(0,1)$. The vector of slope coefficient β is set equal to a vector of ones. In this simulation study, we consider panel datasets of (T=10 and 20) and (n=10,50,100) representing small, medium and large samples respectively, the number of explanatory variables (*p*) is 3 and 4. Due to space constraints, results for *p*=4 is not included. However, the results are consistent with *p*=3. We consider this generated data set as good observations or clean non-correlated data set.

Data is contaminated randomly over all observations (random contamination). The contaminations are done in x-direction at different level of contamination ($\alpha = 0.05, 0.10$). Since contamination is done randomly in the x-direction, the resultant high leverage points may be good or bad leverage points. Also, the magnitude of contamination (MC) was chosen equals to 100 following the idea of Bagheri and Midi (2011). Since we want to focus on collinearity-enhancing observations, a non-correlated data was generated with different percentage of high leverage points. To create outliers in the data, we randomly replaced $100(\alpha)$ percent of good observations of x_1 and x_2 with values equal 100. The simulation was run 1000 times ($R = 1000$).

Table 1 exhibits the VIF values for non-correlated data without HLPs. It is very interesting to observe that for non-correlated data, all VIFs values show no multicollinearity problems (VIF < 5.0). While, it is seen in Tables (2-3) that the CVIF diagnose the problem of multicollinearity in the data in the presence of HLPs, but other VIFs methods do not show multicollinearity regardless of size of samples, and level of contamination. We refer to this situation as high leverage collinearity enhancing observations where the CVIF shows no multicollinearity and show otherwise in the absence and presence of HLPs in the data set, respectively.

It is important to note that when multicollinearity is caused by HLPs (HLCEO), the commonly used methods such as ridge regression, latent root regression, jackknife ridge regression are not the possible options to remedy this problem. Bagheri and Midi (2012) suggested to use Robust Generalized-M (GM) estimator. However, this is not the focus of this paper.

Table 1. VIF values for non-correlated data set (clean data)

n	T	CVIF	RVIF-WMM	RVIF-WGM(DRGP)	RVIF (WGM-FIMGT)
10	10	1.0213	1.0211	1.0640	1.0268
		1.0216	1.0215	1.0631	1.0273
		1.0218	1.0215	1.0597	1.0265
	20	1.0105	1.0107	1.0367	1.0134
		1.0108	1.0109	1.0350	1.0138
		1.0103	1.0105	1.0364	1.0129
50	10	1.0044	1.0044	1.0168	1.0051
		1.0045	1.0043	1.0169	1.0053
		1.0046	1.0044	1.0168	1.0051
	20	1.0021	1.0021	1.0101	1.0028
		1.0019	1.0020	1.0095	1.0026
		1.0019	1.0019	1.0092	1.0026
100	10	1.0023	1.002	1.0093	1.0026
		1.0022	1.002	1.0096	1.0024
		1.0022	1.0021	1.0092	1.0026
	20	1.0009	1.0010	1.0049	1.0012
		1.0010	1.0010	1.0052	1.0013
		1.0010	1.0009	1.0050	1.0011

Table 2. VIF values for non-correlated data with HLPs (MC=100, α=5%)

n	T	CVIF	RVIF-WMM	RVIF-WGM(DRGP)	RVIF (WGM-FIMGT)
10	10	283.68	1.0146	1.0621	1.0265
		283.65	1.0155	1.0639	1.0259
		1.0220	0.9643	1.0637	1.0200
	20	269.26	1.0071	1.0381	1.0139
		269.26	1.0063	1.0368	1.0146
		1.0107	0.9660	1.0387	1.0136
50	10	277.11	1.0019	1.0166	1.0052
		277.11	1.00266	1.0171	1.0051
		1.0041	0.9821	1.0153	1.0049
	20	263.62	1.0006	1.0100	1.0026
		263.62	1.0009	1.0099	1.0027
		1.0019	0.9915	1.0095	1.0026
100	10	55.1253	0.3129	0.5201	0.4881
		55.2341	0.3982	0.5001	0.4082
		0.25231	0.3012	0.4201	0.4001
	20	52.27	0.1987	0.2001	0.1984
		52.27	0.1989	0.2001	0.1996
		0.1992	0.1985	0.2002	0.2011

Table 3. VIF values for non-correlated data with HLPs (MC=100, α=10%)

n	T	CVIF	RVIF-MM	RVIF-GM(DRGP)	RVIF (GM-FIMGT)
10	10	557.47	1.0284	1.0612	1.0274
		557.48	1.0576	1.0576	1.0234
		1.0211	0.9864	1.0602	1.0292
	20	536.43	1.0234	1.0378	1.0108
		536.44	1.0242	1.0387	1.0131
		1.0108	0.9858	1.0366	1.0129
50	10	531.59	1.0106	1.0160	1.0052
		531.59	1.0108	1.0178	1.0041
		1.0038	0.9877	1.0142	1.0047
	20	526.24	1.0112	1.0116	1.0024
		526.24	1.0102	1.0110	1.0016
		1.0019	0.9963	1.0096	1.0027
100	10	55.02	0.3011	0.5011	0.4222
		54.12	0.3002	0.5010	0.4112
		0.2025	0.2963	0.4096	0.3011
	20	52.101	0.1526	0.0171	0.9921
		52.122	0.1782	1.0153	0.9912
		0.1818	0.1880	1.0101	1.0001

6. Numerical examples

In this section, an artificial data and one real data set are used to assess the performance of our proposed methods compared to CVIF measures. Let us first focus on the artificial data set. Non-correlated artificial panel data set with four predictor variables $(\mathrm{p} = 4)$ and sample size (n=10; and T=10) were generated following the same process as the simulation study. To change the pattern of correlation in the data, we modify the data by adding some high leverage points. The modification of data is done by replacing randomly 10% and 5% (the values in parentheses is for 5% HLPs) observations of clean data for the first and second predictor variables $(\mathrm{x}_1 \text{ and } \mathrm{x}_2)$by arbitrary large values (100).

Figure 1(a) shows the boxplot of the modified data set which has some outliers, and Figure 1 (b) shows that x_1 and x_2 are highly correlated ($\rho_{12} = 0.98$). The results of the classical and robust VIF methods are presented in Table 4. The results indicate that all the VIF methods correctly detect no multicollinearity problem in the original data set (no HLPs and no multicollinearity). For the modified data set, the CVIF indicates that there is multicollinearity problem in the data. However, robust VIFs diagnose that there is no multicollinearity.

Table 4. The VIF values for the classical and robust diagnostic methods for the original (non-correlated) and modified artificial data set n=10;t=10; MC=100 in panel data

Variables	method	Original Artificial data set				Modified Artificial data set 5%(10%)			
		CVIF	RVIF-WMM	RVIF (WGM (DRGP))	RVIF (WGM-FIMGT)	CVIF	RVIF-MM	RVIF (WGM (DRGP))	RVIF (WGM-FIMGT)
x_1	MM-CENTERING	1.0250	1.0257	1.0666	1.0395	287.80 (550.21)	1.0256 (1.0168)	1.0733 (1.0590)	1.0410 (1.0364)
x_2		1.0039	1.0075	1.0309	1.0024	288.04 (550.36)	1.0087 (1.0029)	1.0476 (1.0342)	1.0073 (1.0046)
x_3		1.0237	1.0120	1.0508	1.0254	1.018 (1.018)	1.0078 (1.0076)	1.0507 (1.0545)	1.0198 (1.0209)
x_4		1.0022	1.0023	1.0226	1.0021	1.018 (1.025)	1.0230 (1.0343)	1.0230 (1.0273)	1.02873 (1.0506)

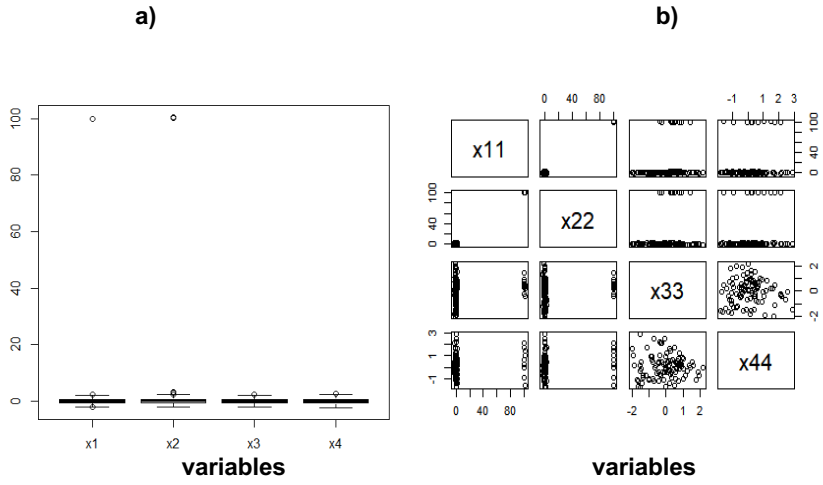

Figure 1. The boxplot (a) and scatter plot (b) for modified correlated artificial data set after MM-centering

An airline firms data taken from Greene(2007) is the first real example in this paper. The data represents the relationship between the response variable (cost) and three predictor variables (output, fuel price, and load factor) over 15 yearly observations (1970-1984). This data set has no outlier. To see the effect of outlier, the first good observation of the original data for x1 and x2 are replaced with a fixed value equals to 100. According to Greene (2007), this model measured the output in "revenue passenger miles". The load factor is defined as the average rate at which seat's on the airline's planes are filled. A straight forward equation can be formed as an illustration where the total cost of production is fitted into a multiple linear regression model:

$$lncos\, t_{it} = \alpha_i + ln\, o\, utput_{it} + ln\, f\, uelprice_{it} + loadfactor_{it} + \varepsilon_{it}$$

The CVIF, RVIF(WMM) , RVIF(WGM(DRGP)) and RVIF(WGM -FIMGT) were then applied to the data. Table 5 exhibits the results for clean data (original) and contaminated data (modified). The results indicate that all the VIF methods detect no multicollinearity problem in original data set (no HLPs and no multicollinearity). For the modified data set, the CVIF indicates that there is multicollinearity problem in

the data. However, the three proposed robust VIFs diagnose that there is no multicollinearity. The results of the numerical examples are consistent with the results of the simulation study.

Table 5. The VIF values for the classical and robust diagnostic methods for the airline data set.

Variables	method	Original data set				Modified data set			
		CVIF	RVIF-WMM	RVIF (WGM (DRGP))	RVIF (WGM-FIMGT)	CVIF	RVIF-MM	RVIF (WGM (DRGP))	RVIF (WGM-FIMGT)
$\hat{\beta}_1$	MM-	3.74829	2.8660	4.093620	4.04705	432.9652	2.747681	4.132489	4.044109
$\hat{\beta}_2$		3.66192	2.8339	4.337642	4.02317	430.2707	2.711077	4.296857	4.058660
$\hat{\beta}_3$		2.09776	1.6758	3.462942	2.54735	.446713	1.594729	3.385232	2.547356

7. Conclusion

The aim of this article is to formulate a reliable robust inflation factor (RVIF) to correctly diagnose multicollinearity in a panel data model in the presence of HLCEO. We have developed robust RVIF(WGM-FIMGT) in this regard based on our newly developed WGM-FIMGT estimator. We also extent the robust RVIF-MM and robust RVIF-DRGP of Bagheri and Midi (2011) in multiple linear regression to fixed effect panel data model. In the absence of HLPs, all VIFs methods show no multicollinearity. Nevertheless, all methods still show no multicollinearity in the presence of HLPs except the CVIF. This situation is referred to as HLCEO. In the presence of HLCEO, the CVIF wrongly diagnosed multicollinearity but all the RVIFs are very successful to show otherwise. In this situation, relying on CVIF, statistics practitioners will simply employ the commonly used methods such as latent root regression and ridge regression estimator to rectify this problem. This will produce inefficient estimates and will lead to misleading conclusions. The correct way to deal with this problem is by using robust Generalized-M estimator. Although, the performance of RVIF-MM, RVIF(WGM-DRGP) and RVIF(WGM-FIMGT) give similar conclusion, the main attraction of RVIF(WGM-FIMGT) is that its computation running times is much shorter than the other methods as shown by Lim and Midi (2016) since it employs the ISE which is very fast algorithm to compute the robust location and covariance matrix .

References

[1] Kamruzzaman, M. and Imon, A.H.M.R.. High leverage point: another source of multicollinearity. Pakistan Journal of Statistics. 18(3): 435-448 ,2002.

[2] Bagheri, A., and Midi, H. On the performance of robust variance inflation factors. International Journal of Agricultural and Statistical Sciences. 7(1): 31-45, 2011.

[3] Bagheri, A., Habshah M. and Imon, A.H.M.R. A novel collinearity-influential observation diagnostic measure based on a group deletion approach. Communications in Statistics - Simulation and Computation. 41(8): 1379-1396, 2012.

[4] Midi, H., Bagheri, A. and Imon, A.H.M.R. A Monte Carlo simulation study on high leverage collinearity-enhancing observation and its effect on multicollinearity pattern. Sains Malaysiana. 40(12), 1437-1447, 2011.

[5] Midi, H., Ismaeel S. S., and Arasan J. On the performance of fast Robust Variance Inflation factor based on index set equality. Journal of Engineering and Applied Sciences. 13(16): 6634-6638 , 2018.
[6] Habshah, M., Norazan, M.R., and Imon, A.H.M.R. The performance of Diagnostic-Robust Generalized Potentials for the identification of multiple high leverage points in linear regression. Journal of Applied Statistics. 36(5): 507-520, 2009.
[7] Hadi, A.S. Diagnosing collinearity-influential observations. Computational Statistics and Data Analysis. 1988.7(2): 143-159 ,2011.
[8] Sengupta, D. and Bhimasankaram P. On the roles of observations in collinearity in the linear model. Journal of the American Statistical Association. 92(439): 1024-1032 ,1997.
[9] Bramati, M.C. and Croux C. Robust estimators for the fixed effects panel data model. The Econometrics Journal. 10(3): 521-540, 2007.
[10] Baltagi, B. Econometric analysis of panel data. John Wiley & Sons, 2008.
[11] Bakar, N. M. A. and H. Midi. Robust centering in the fixed effect panel data model. Pakistan Journal of Statistics. 31(1):33-48, 2015.
[12] Bagheri A, and Midi H. Diagnostic plot for the identification of high leverage collinearity-influential observations. Statistics and Operations Research Transactions. 39(1): 51-70, 2015.
[13] Ismaeel, S.S., and Midi, H. Robust within group estimator for fixed effect panel data. Pakistan Journal of Statistics. 34(4): 297-310, 2018.
[14] Lim, H. A. and H. Midi. Diagnostic Robust Generalized Potential Based on Index Set Equality (DRGP (ISE)) for the identification of high leverage points in linear model. Computational Statistics. 3(31): 859-877, 2016.
[15] Rousseeuw, P.J., and Van Zomeren, B.C. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association. 85(411): 633-639, 1990.
[16] Maronna, R.A., Martin, R.D., and Yohai, V.J. Robust Statistics Theory and Methods. New York: Wiley and Sons, 2006.
[17] Krasker, W. S. and Welsch R.E. Efficient bounded-influence regression estimation. Journal of the American Statistical Association. 77(379): 595-604, 1982
[18] Rousseeuw, P.J., and Leroy, A.M. Robust Regression and Outlier Detection. New York: Wiley, 1987.
[19] Coakley, C.W., and Hettmansperger, T.P. A bounded influence, high breakdown, efficient regression estimator. Journal of the American Statistical Association. 88(423): 872-880, 1993.
[20] Midi, H., Sani, M., Ismaeel, S.S., and Arasan, J. Fast Improvised Influential Distance for the Identification of Influential Observations in Multiple Linear Regression. Sains Malaysiana. *50*(7), 2085-2094, 2021.
[21] Rousseeuw, P.J. Least median of squares regression. Journal of the American Statistical Association. 79(388): 871-880,1984.
[22] Greene, W.H. Econometric Analysis. 6th Edition, Prentice Hall, 2007.