Journal of Engineering and Applied Sciences 13 (16): 6634-6638, 2018
ISSN: 1816-949X

On the Performance of Fast Robust Variance Inflation Factor Based on Index Set Equality

[1, 3]Habshah Midi, [1, 2]Shelan Saied Ismaeel and [1]Jayanthi Arasan
[1]Faculty of Science and Institute for Mathematical Research,
Universiti Putra Malaysia 43400, UPM Serdang, Selangor, Malaysia
[2]Department of Mathematics, Faculty of Science, University of Zakho, Zakho, Iraq
[3]Applied and Computational Statistics Laboratory, Institute for Mathematical Research,
Serdang, Malaysia

Abstract: The detection of multicollinearity is very crucial, so that, proper remedial measures can be taken up in their presence. The widely used diagnostic method to detect multicollinearity in multiple linear regressions is by using Classical Variance Inflation Factor (CVIF). It is now evident that the CVIF failed to correctly detect multicollinearity when high leverage points are present in a set of data. Robust Variance Inflation Factor (RVIF) has been introduced to remedy this problem. Nonetheless, the computation of RVIF takes longer time because it is based on robust GM (DRGP) estimator which depends on Minimum Volume Ellipsoid (MVE) estimator that involves a lot of computer times. In this study, we propose a fast RVIF (FRVIF) which take less computing time. The results of the simulation study and numerical examples indicate that our proposed FRVIF successfully detect multicollinearity problem with faster rate compared to other methods.

Key words: Generalized-M, high leverage points, robust variance inflation factor, multicollinearit, estimator, computer

INTRODUCTION

One of the assumptions of the general linear regression model is that there is no correlation (or no multicollinearity) between the explanatory variables. When this assumption is not met, the ordinary least squares estimates may have wrong sign problem have large variances and this would lead to erroneous interpretation. It arises when there is a near-liner dependency among explanatory variables (x-direction) in multiple linear regression models. It may also result due to the data collection method employed, constrains on the model, model specification and over determined model.

CVIF is the commonly used diagnostic method for detecting multicollinearity in linear regression. It measures how much the variances of the estimated regression coefficients are inflated as compared to when the predictors are not correlated (Belsley, 1991; Belsley *et al*., 1980; Stine, 1995). It has done well in a clean data set but its performance becomes poor in the presence of high leverage points (Midi *et al*., 2010; Bagheri and Midi, 2009; Bagheri, 2011) has shown that the CVIF cannot detect multicollinearity when high leverage points are present in a data set. They have developed two robust VIFs namely the VIF which is based on MM and the VIF which is based on GM (DRGP) which they called them RVIF(MM) and RVIF (GM(DRGP)), respectively.

The RVIF (MM) which is based on MM estimator moderately identifies multicollinearity but failed to detect multicollinearity when high leverage points are present. The RVIF (GM (DRGP)) method which is based on DRGP able to detect multicollinearity in the absence and presence of high leverage points. However, the RVIF (GM (DRGP)) takes longer computational time as it is based on Minimum Volume Ellipsoid (MVE) which has slow convergent rate in the computation of robust Mahalanobis distance (Rousseeuw and Leroy, 2005). Their research has motivated us to propose an improvised RVIF which is relatively faster than the RVIF (GM (DRGP)).

Multicollinearity diagnostic measures: A simple technique for revealing multicollinearity issue is by checking the simple correlations between predictors. A high value of correlation coefficient indicates the existence of serious problem of collinearity.

When there are more than two independent variables, the simple correlation may mislead conclusion, even

Corresponding Author: Habshah Midi, Faculty of Science and Institute for Mathematical Research, Universiti Putra Malaysia 43400, UPM Serdang, Selangor, Malaysia

if they are all very low, they could hide the serious multicollinearity problems. This happen if there is no clear overlapping among predictors but they have a cumulative effect (Montgomery *et al*., 2001; Kutner *et al*., 2004; Freund *et al*., 2006).

Classical variance inflation factor: Marquardt (1970) developed a diagnostics method which is known as Variance Inflation Factor (CVIF) to detect multicollinearity in a data. The CVIF is the most popular method to identify multicollinearity and it is given by:

$$VIF_j = \frac{1}{1-R_j^2},\ j = 1, 2, \ldots, p \quad (1)$$

Where:

R^2_j = The coefficient of multiple determination when
x_j = Regressed on other
$X_{(p-1)}$ = Variables in the model, using the Ordinary Least Squares (OLS) method

In general, if $VIF_{max} \bullet (5,10)$ indicates that there is moderate multicollinearity among all of predictors and when $VIF_{max} \bullet 10$ indicates that there is a severe multicollinearity (Belsley, 1991).

RVIF (MM): The OLS estimates which is used in the computation of CVIF is known to be easily affected by outliers. As such, Bagheri (2011) proposed RVIF (MM) based on the robust MM estimator (Rousseeuw and Leroy, 2005) which is defined as:

$$RVIF_j(MM) = \frac{1}{1\text{-}RR_j^2(MM)},\ j = 1, 2, \ldots, p \quad (2)$$

Where:

RR^2_j = The coefficient of multiple determination when
x_j = Regressed on other
$X_{(p-1)}$ = Variables in the model using MM estimator

RVIF (GM (DRGP)): Since, the MM estimator has no bounded influence property (Bagheri *et al*., 2012) developed another RVIF which is based on generalized M estimator which is robust on both outliers in x and Y directions. They called the developed diagnostic method as RVIF (GM (DRGP)) as it is based on Diagnostic Robust Generalized Potential (DRGP) of Habshah *et al*. (2009), the RVIF (GM (DRGP)) is given by:

$$RVIF_j(GM(DRGP)) = \frac{1}{1\text{-}RR_j^2(GM(DRGP))},\ j = 1, 2, \ldots, p \quad (3)$$

where, RR^2_j is the coefficient of multiple determination when x_j is regressed on other $x_{(p-1)}$ variables in the model using GM (DRGP) estimator. $RR^2_j(GM(DRGP))$ is defined as follows:

$$RR^2\left(GM\left(DRGP\right)\right) = 1 - \frac{\sum_{i=1}^{n} w_i r_i^2}{\sum_{i=1}^{n} w_i \left(y_i - \bar{y}\right)^2} \quad (4)$$

where, w_i and $r_i$are the robust weights and residuals obtained from GM (DRGP), respectively. The $\bar{y}$ is the weighted average of y, given as:

$$\bar{y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \quad (5)$$

Prior to obtaining the $RR^2_j(GM(DRGP))$, the GM (DRGP) needs to be established. The GM (DRGP) is summarized in the following steps:

Step 1: For 'k' is a number of iteration, begin by setting k = 0 and compute the coefficients ($\bullet_j$, j = 0,1, ..., p) and residuals (r_ii, i = 1, 2, ..., n) for S-estimator.

Step 2: For i = 1, 2, ..., n compute initial weight function depend on DRGP as:

$$\pi_i = \min[1, \frac{median(p_{ii}) + 3Mad(p_{ii})}{p_{ii}}]$$

where P_{ii} is DRGP (MVE) of Habshah *et al* (2009).

Step 3: Scale residuals by $\bullet$ which is defined as :

$$\hat{\tau} = 1.4826(1+5(n-p))median\left|r_i\right|$$

Step 4: Define the initial weights as:

$$w_{ik} = \frac{\hat{\tau}_k \times \pi_i}{r_{ik}} \psi(\frac{r_{ik}}{\hat{\tau}_k \times \pi_i}) \quad (6)$$

for i = 1, 2, ..., n where a Huber's $\bullet$ -function is applied.

Step 5: Use these weights to obtain a weighted least squares estimates.

Step 6: Repeat steps 3-5 until convergence. That is, iterate until the change in the estimated parameters is small.

FAST RVIF (GM(DRGP)): The RVIF (GM(DRGP)) is known to be able to detect multicollinearity problem in the presence of high leverage points. The weakness of this method is that the computation of the DRGP in the second step of GM (DRGP) takes longer computational times as it is based on the Minimum Volume Ellipsoid (MVE). In this situation, Lim and Midi (2016)improvised the DRGP by using Index Set Inequality (ISE) instead of using the MVE in the first step of the computation of DRGP. With this modification, it has been shown that the DRGP has taken less computational time. In order to propose fast RVIF (GM (DRGP)) , we adapt the improvised DRGP by Lim and Midi (2016) to compute RR^2_j(GM(DRGP)). The DRGP (ISE) method can be summarized as follows:

Step 1: Compute the Robust Mahalanobis Distance (RMD_i) for each ith point, using Index Set Inequality (ISE) by Salleh (2013).

Step 2: Any observation in which its RMD_i exceeds the cut-off value, i.e, RMD_i>Median (RMD_i)+Mad (RMD_i), they are considered as suspected HLPs and be included in the deletion D group, the remaining cases are included in the R group.

Step 3: Compute the P_{ii} based on the above D and R sets as follows:

$$p_{ii} = \begin{cases} h_{ii}^{(-D)} & \text{for } i \in D \\ \dfrac{h_{ii}^{(-D)}}{1-h_{ii}^{(-D)}} & \text{for } i \in R \end{cases} \quad (7)$$

Where:

$$h_{ii}^{(-D)} = x_i^t (X_R^t X_R)^{-1} x_i \quad i = 1,2, \ldots, n \quad (8)$$

Step 4: For all the set D with p_{ii}>Median (p_{ii})+3Q_n(p_{ii}) are declared as HLPs, else, the case with least P_{ii} will be put back into set R and repeat step 3 and 4 until all p_{ii}>Median (p_{ii})+3Q_n(p_{ii}).

$Q_n = c\{|x_i - x_j|; <j\}_{(K)}$ is a pairwise order statistic of all distance proposed by Rousseeuw and Leroy (2005) where $k = {}^hC_2 \bullet {}^nC_{2/4}$ and h = [n/2]+1. The used of c = 2.2219 as this value will provide Q_n a consistent estimator for Gaussian data.

In the simulation study conducted by Lim and Midi (2016), they have shown that the running time of the ISE is much faster than the MVE and MCD. Hence by incorporating the ISE in the GM (DRGP) algorithm will subsequently decrease the running time of the FRVIF method.

MATERIALS AND METHODS

Monte-Carlo simulation study: A Monte-Carlo simulation study is employed in order to assess the performance of fast RVIF. We consider the multivariate linear regression model as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \quad (9)$$

where, • is distributed as N (0,1). The predictor variables were generated followed the Lawrence and Arthur procedure which is defined as:

$$x_{ij} = \rho v_{i4} + (1-\rho^2)^{1/2} v_{ij},\ i = 1,2, \cdots, n;\ j = 1,2 \text{ and } 3 \quad (10)$$

The correlation coefficient (•) was chosen to be very high at 0.98. We consider samples (n = 20, 50, 100, 200 and 300) and different level of contamination (• = 0.05, 0.10, 0.15 and 0.20). Moreover, the magnitude of contamination (MC) was chosen equals to 100 following the idea of Mohammed and Midi. To add high leverage point to data, the first 100 (•/2)% of observations for x1 and the last 100(•/2)% of observations for x2 have been replaced by different magnitude of contaminations values.

We run the simulation 1000 times for consistency. Table 1-3 exhibit the VIF values for correlated data without HLPs and with HLPs, respectively. It can be observed from Table 1 that the CVIF and FRVIF (GM(DRGP)) can correctly identify the problem of multicollinearity except for the RVIF (MM).

It is interesting to observe the behavior of the CVIF and RVIF (MM) in the presence of high leverage points Table 2 and 3. In this situation, the CVIF and RVIF (MM) cannot detect multicollinearity problem in the data. On the other hand, the RVIF (MG(DRGP)) still can correctly revealed the multicollinearity problem, irrespective of the percentage of high leverage points and sample size.

Table 1: VIF values for correlated data with no HLP (• = 0%)

n	CVIF	RVIF-MM	FRVIF- GM (DRGP)
	21.79977	4.110346	30.591950
20	22.09634	4.152511	31.142910
	21.31794	4.116280	30.085000
	18.15998	3.994056	33.379080
50	18.31078	4.042969	32.993010
	18.17335	4.030211	32.913880
	17.61057	4.104120	36.706700
100	17.66150	4.114590	37.025000
	17.64393	4.049040	37.410900
200	17.37899	4.122235	41.476730
	17.46795	4.116485	41.735420
	17.41924	4.130874	41.567910
300	17.35104	4.125675	44.364850
	17.39302	4.135690	44.768620
	17.26760	4.124726	4.124726

Table 2:-VIF values for correlated data with HLP (MC=100, • = 5%, • = 10%)

	5%			10%		
n	CVIF	RVIF-MM	FRVIF- GM(DRGP)	CVIF	RVIF-MM	FRVIF- GM(DRGP)
20	1.0700	3.6000	3.71740	1.06910	3.2011	1.6472
	1.0963	3.6310	3.60160	1.07560	3.2224	1.6763
	1.1588	3.5870	29.38830	1.13910	3.2070	29.0250
50	1.0258	3.5993	3.40927	1.02590	3.2766	2.1913
	1.0288	3.5986	3.45955	1.02720	3.2681	2.2041
	1.0541	3.5377	32.81100	1.04990	3.2431	32.0250
100	1.0169	3.6459	3.86407	1.01600	3.1331	2.0868
	1.0155	3.6585	3.87577	1.07560	3.0931	2.0880
	1.0319	3.6003	37.13580	1.04990	3.0409	35.9800
200	1.0094	3.5581	3.62317	1.00970	3.1322	2.2540
	1.0099	3.5556	3.65043	1.00990	3.0952	2.2670
	1.0182	3.5117	40.80650	1.01420	3.0466	39.9100
300	1.0080	3.6024	4.04590	0.79550	2.4636	1.8630
	1.0077	3.6036	4.05050	0.79537	2.4434	1.8680
	1.0148	3.5617	43.12670	0.79762	2.4058	33.3700

Table 3: VIF values for correlated data with HLP (MC = 100, • = 15%, • = 20%)

	15%			20%		
N	CVIF	RVIF-MM	FRVIF-GM(DRGP)	CVIF	RVIF-MM	FRVIF- GM(DRGP)
20	1.06919	3.201100	1.647210	1.08074	2.7912	1.2471
	1.07560	3.222480	1.676380	1.08424	2.7939	1.2539
	1.13910	3.207190	29.025100	1.13863	2.8274	27.9840
50	1.02570	3.017690	1.740030	1.03349	2.7648	1.3706
	1.02710	3.040870	1.737820	1.03538	2.7634	1.3694
	1.04470	3.023320	31.390100	1.04361	2.7519	29.8190
100	1.01750	2.849630	1.728319	1.02491	2.6052	1.4504
	1.01710	2.809690	1.726987	1.02377	2.5236	1.4511
	1.02349	2.765900	35.228370	1.02377	2.5403	33.8590
200	1.01318	2.769357	1.774541	1.01833	2.5866	1.5281
	1.01316	2.737710	1.770475	1.01839	2.5113	1.5338
	1.01320	2.694498	38.765800	1.01183	2.5356	37.8780
300	1.01110	2.782220	1.864912	1.01665	2.6044	1.5739
	1.01080	2.746160	1.866877	1.01683	2.4840	1.5752
	1.00954	2.706560	41.314660	1.01683	2.5149	40.3860

RESULTS AND DISCUSSION

Numerical examples: Body fat dataset is used to evaluate the performance of our proposed method. This data set contains 20 observations and has three predictors (p = 3). Kutner *et al.* (2004) showed that this dataset has multicollinearity problem. In order to see the effect of HLPs on the VIF measures, we replaced 5 and 10% of the good observations for x1 with 100 to create high leverage points in the data. Table 4 presents the coefficient of determination (R^2) and VIF for the original dataset. The results of R^2 and VIF for all diagnostic measures except RVIF (MM) indicate a high correlation among the predictor variables and showed that this dataset has multicollinearity problem.

The results of Table 5 signify that the CVIF failed to identify the multicollinearity in the dataset while the RVIF (MM) identifies that there is moderate multicollinearity. On the other hand, the RVIF (GM(DRGP)) successfully identify a serious multicollinearity problem in the dataset.

Table 4: The R^2 and VIF values for the original body fat data set

	CVIF		RVIF-MM		FRVIF(GM (DRGP))	
Variables	R^2	VIF	R^2	RVIF	R^2	RVIF
X_1	0.9985	708.842	0.84017	6.2567	0.998687	762.138
X_2	0.9982	564.343	0.84465	6.4373	0.998402	625.990
X_3	0.9904	104.606	0.80516	5.1326	0.989449	94.7839

Table 5: R^2 and VIF values for modified body fat data set

	CVIF		RVIF-MM		FRVIF (GM(DRGP))	
Variables	R^2	VIF	R^2	RVIF	R^2	RVIF
X_1	0.05984	1.06365	0.80502	5.12875	0.91970	12.4534
	(0.0705)	(1.0758)	(0.8437)	(6.3988)	(0.9563)	(22.895)
X_2	0.02817	1.02899	0.79511	4.88079	0.9966	299.145
	(0.0247)	(1.0253)	(0.8351)	(6.0655)	(0.9979)	(498.44)
X_3	0.05171	1.05454	0.77367	4.41845	0.9923	129.932
	(0.0660)	(1.0707)	(0.8198)	(5.5500)	(0.9907)	(107.76)

CONCLUSION

The commonly used CVIF method is very successful in detecting multicollinearity problem in a data set.

However, it failed to diagnose multicollinearity problem in the presence of high leverage points. The performance of RVIF (MM) is not good for both situations. In this regard, we propose fast robust RVIF method for detecting multicollinearity in a data set. The proposed method is formulated by incorporating fast DRGP of Lim and Midi (2016). The results of our study show that our proposed fast RVIF (GM(DRGP)) can detect multicollinearity irrespective of whether high leverage points are present in a data set. Hence, we suggest using this method to get correct interpretation of a data, so that, proper remedial measure can be taken up.

REFERENCES

Alguraibawi, M., H. Midi and A.H.M.R. Imon, 2015. A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model. Math. Prob. Eng., 2015: 1-12.

Bagheri, A. and H. Midi, 2009. Robust estimations as a remedy for multicollinearity caused by multiple high leverage points. J. Math. Statist., 5: 311-321.

Bagheri, A., 2011. Robust estimation methods and robust multicollinearity diagnostics for multiple regression model in the presence of high leverage collinearity-influential observations. Ph.D Thesis, Universiti Putra Malaysia, Seri Kembangan, Malaysia.

Bagheri, A., M. Habshah and R.H.M.R. Imon, 2012. A novel collinearity-influential observation diagnostic measure based on a group deletion approach. Commun. Stat.-Simulat. Comput., 41: 1379-1396.

Belsley, D.A., 1991. Conditioning Diagnostics: Collinearity and Weak Data in Regression. John Wiley and Sons, Hoboken, New Jersey, USA., ISBN:9780471528890, Pages: 396.

Belsley, D.A., E. Kuh and R.E. Welsch, 1980. Regression Diagnostics: Identifying Influential Data and Sources of Colinearity. John Willey and Sons Inc., New York.

Freund, R.J., J.W. William and S. Ping, 2006. Regression Analysis. Academic Press, Cambridge, Massachusetts, USA., ISBN-13:978-0-12-088597-8, Pages: 459.

Habshah, M., M.R. Norazan and A.H.M.R. Imon, 2009. The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. J. Applied Statist., 36: 507-520.

Kutner, M.H., C. Nachtsheim and J. Neter, 2004. Applied Linear Regression Models. 4th Edn., McGraw-Hill/Irwin, New York, USA., ISBN:9780073013442, Pages: 701.

Lim, H.A. and H. Midi, 2016. Diagnostic robust generalized potential based on index set equality (DRGP (ISE)) for the identification of high leverage points in linear model. Comput. Stat., 31: 859-877.

Marquardt, D.W., 1970. Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. Technometrics, 12: 591-612.

Midi, H., A. Bagheri and A.H.M.R. Imon, 2010. The application of robust multicollinearity diagnostic method based on robust coefficient determination to a non-collinear data. J. Applied Sci., 10: 611-619.

Montgomery, E., M.P. Bronner, J.R. Goldblum, J.K. Greenson and M.M. Haber *et al.*, 2001. Reproducibility of the diagnosis of dysplasia in Barrett esophagus: A reaffirmation. Hum. Pathol., 32: 368-378.

Rousseeuw, P.J. and A.M. Leroy, 2005. Robust Regression and Outlier Detection. Vol. 589, John Wiley & Sons, Hoboken, New Jersey, USA., Pages: 331.

Salleh, R.M., 2013. A robust estimation method of location and scale with application in monitoring process variability. Ph.D Thesis, Universiti Teknologi Malaysia, Johor Bahru, Malaysia.

Stine, R.A., 1995. Graphical interpretation of variance inflation factors. Am. Stat., 49: 53-56.