The Performance of fast robust Variance Inflation Factor in the presence of high leverage points

Habshah Midi\textsuperscript{1,3}, Shelan Saied Ismaeel\textsuperscript{1,2}, Jayanthi Arasan\textsuperscript{1}

\textsuperscript{1}Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
\textsuperscript{2}Department of Mathematics, Faculty of Science, University of Zako, Iraq,
\textsuperscript{3}Applied & Computational Statistics Laboratory Institute for Mathematical Research,

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 regression 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 paper, 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.

Keywords: Generalized –M, High Leverage Points, Robust Variance Inflation Factor, and Multicollinearity.

1. 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 linear 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, 1-3] It has done well in a clean data set but its performance becomes poor in the presence of high leverage points (see Habshah e. al. [4-6]) 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 identify 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 et al.[7]). Their work has motivated us to propose an improvised RVIF which is relatively faster than the RVIF (GM (DRGP)).

This chapter is organized as follows: the commonly used Variance Inflation Factor and the proposed FRVIF are presented in Section 2. Section 3 and Section 4 discuss the Monte Carlo simulation study and numerical example, respectively to assess the performance of our propose RVIF method. Finally, some concluding remarks are given in Section 5.

2. 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 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 (for more details one can refer [8-10].

2.1 Classical Variance Inflation Factor

Marquardt [11] 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:
is the coefficient of multiple determination when \( x_j \) is regressed on other \( X_{(p-1)} \) variables in the model. Using the Ordinary Least Squares (OLS) method, in general, if \( VIF_{\text{max}} \in (5, 10) \) indicates that there is moderate multicollinearity among all of predictors, and when \( VIF_{\text{max}} \geq 10 \) indicates that there is a severe multicollinearity (Belsley et al. [1]).

2.2 RVIF(MM)

The OLS estimates which is used in the computation of CVIF is known to be easily affected by outliers. As such, Bagheri et al. [6] proposed RVIF(MM) based on the robust MM estimator (Rousseeuw et al. [7]) which is defined as:

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

where \( RR_j^{2}(MM) \) is the coefficient of multiple determination when \( x_j \) is regressed on other \( X_{(p-1)} \) variables in the model using MM estimator.

2.3 RVIF(GM(DRGP))

Since the MM estimator has no bounded influence property, Bagheri et al. [12] 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. [13]. The RVIF (GM(DRGP)) is given by:

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

where \( RR_j^{2}(GM(DRGP)) \) is the coefficient of multiple determination when \( x_j \) is regressed on other \( X_{(p-1)} \) variables in the model using GM (DRGP) estimator.

\[
RR_j^{2}(GM(DRGP)) = 1 - \frac{\sum_{i=1}^{n} w_i r_{ij}^2}{\sum_{i=1}^{n} w_i (y_i - \bar{y})^2}
\]

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

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

Prior to obtaining the \( RR_j^{2}(GM(DRGP)) \), 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 \( (\hat{\beta}_j, j = 0, 1, \ldots, p) \) and residuals \( (\hat{r}_i, i = 1, 2, \ldots, n) \) for S-estimator.

Step 2: For \( i = 1, 2, \ldots, n \) compute initial weight function \( \pi_i \) depend on DRGP as

\[
\pi_i = \min[1, \frac{\text{median}(p_u) + 3\text{Mad}(p_u)}{p_u}]
\]

where \( p_u \) is DRGP(MVE) of Habshah et al. [13].

Step 3: Scale residuals by \( \hat{r} \) which is defined as

\[
\hat{r} = 1.4826(1 + 5(n - p))\text{median}|r|
\]

Step 4: Define the initial weights as

\[
w_{ik} = \frac{\hat{r}_k \times \pi_i}{r_{ik} \times \hat{r}_k \times \pi_i}
\]

for \( i = 1, 2, \ldots, n \), where a Huber’s \( \psi \)-function is applied.

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

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

2.4 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 Habshah [14] 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 of Lim and Habshah [14] to compute \( RR_j^{2}(GM(DRGP)) \).

The DRGP (ISE) method can be summarized as follows:

Step 1: Compute the Robust Mahalanobis Distance (RMD) for each \( i \) point, using Index Set Inequality (ISE) of Rohayu [15].

Step 2: Any observation in which its RMD exceeds the cut-off value, i.e.,

\[
RMD_i > \text{Median}(RMD_i) + \text{Mad}(RMD_i)
\]

are considered as suspected influential and be included in the deletion D group, the remaining cases are included in the R group.

Step 3: Compute the \( p_i \) based on the above D and R sets as follows:

\[
p_i = \begin{cases} 
    h_i^{(D)} & \text{for } i \in D \\
    h_i^{(R)} & \text{for } i \in R
\end{cases}
\]
3. 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 \( \varepsilon \) is distributed as \( N(0, I) \). The predictor variables were generated followed the Lawrence and Arthur (1990) procedure, which is defined as

\[ x_{ij} = \rho v_{ij} + (1 - \rho^2)^{1/2} v_y, \quad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, 3 \quad (10) \]

The correlation coefficient (\( \rho \)) was chosen to be very high at 0.98. We consider samples \( n = 20, 50, 100, 200 \) and 300) and different level of contamination(\( \alpha = 0.05, 0.10, 0.15 \) and 0.20). Moreover, the magnitude of contamination (MC) was chosen equals to 100 following the idea of Mohammed and Midi [16]. To add high leverage point to data, the first 100(\( \alpha/2 \)) percent of observations for \( x_1 \) and the last 100(\( \alpha/2 \)) percent of observations for \( x_2 \) have been replaced by different magnitude of contaminations values.

We run the simulation 1000 times for consistency. Table 1 and Tables 2-3 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(2-3). 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>
<thead>
<tr>
<th>( n )</th>
<th>CVIF</th>
<th>RVIF-MM</th>
<th>FRVIF-GM(DRGP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>1.0700</td>
<td>3.6000</td>
<td>3.7174</td>
</tr>
<tr>
<td>50</td>
<td>1.0963</td>
<td>3.6310</td>
<td>3.6016</td>
</tr>
<tr>
<td>100</td>
<td>1.1588</td>
<td>3.5873</td>
<td>29.3883</td>
</tr>
<tr>
<td>200</td>
<td>1.0258</td>
<td>3.5993</td>
<td>3.40927</td>
</tr>
<tr>
<td>300</td>
<td>1.0541</td>
<td>3.5377</td>
<td>32.8110</td>
</tr>
</tbody>
</table>

Table 1: VIF values for correlated data with no hlp(\( \alpha=0\% \))

<table>
<thead>
<tr>
<th>( n )</th>
<th>CVIF</th>
<th>RVIF-MM</th>
<th>FRVIF-GM(DRGP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>1.0700</td>
<td>3.6000</td>
<td>3.7174</td>
</tr>
<tr>
<td>50</td>
<td>1.0963</td>
<td>3.6310</td>
<td>3.6016</td>
</tr>
<tr>
<td>100</td>
<td>1.1588</td>
<td>3.5873</td>
<td>29.3883</td>
</tr>
<tr>
<td>200</td>
<td>1.0258</td>
<td>3.5993</td>
<td>3.40927</td>
</tr>
<tr>
<td>300</td>
<td>1.0541</td>
<td>3.5377</td>
<td>32.8110</td>
</tr>
</tbody>
</table>

Table 2: VIF values for correlated data with hlp (MC=100, \( \alpha=5\% \alpha=10\% \))
Table 3: VIF values for correlated data with hlp (MC=100, α=15% α=20%)

<table>
<thead>
<tr>
<th>n</th>
<th>CVIF</th>
<th>RVIF-MM</th>
<th>FRVIF-GM(DRGP)</th>
<th>CVIF</th>
<th>RVIF-MM</th>
<th>FRVIF-GM(DRGP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>1.06919</td>
<td>3.20110</td>
<td>1.64721</td>
<td>1.08074</td>
<td>2.7912</td>
<td>1.2471</td>
</tr>
<tr>
<td></td>
<td>1.0756</td>
<td>3.22248</td>
<td>1.67638</td>
<td>1.08424</td>
<td>2.7939</td>
<td>1.2539</td>
</tr>
<tr>
<td></td>
<td>1.1391</td>
<td>3.20719</td>
<td>29.0251</td>
<td>1.13863</td>
<td>2.8274</td>
<td>27.984</td>
</tr>
<tr>
<td>50</td>
<td>1.0257</td>
<td>3.01769</td>
<td>1.74003</td>
<td>1.03349</td>
<td>2.7648</td>
<td>1.3706</td>
</tr>
<tr>
<td></td>
<td>1.0271</td>
<td>3.04087</td>
<td>1.73782</td>
<td>1.03538</td>
<td>2.7634</td>
<td>1.3694</td>
</tr>
<tr>
<td></td>
<td>1.0447</td>
<td>3.02332</td>
<td>31.3901</td>
<td>1.04361</td>
<td>2.7519</td>
<td>29.819</td>
</tr>
<tr>
<td>100</td>
<td>1.0175</td>
<td>2.84963</td>
<td>1.728319</td>
<td>1.02491</td>
<td>2.6052</td>
<td>1.4504</td>
</tr>
<tr>
<td></td>
<td>1.0171</td>
<td>2.80969</td>
<td>1.726987</td>
<td>1.02377</td>
<td>2.5236</td>
<td>1.4511</td>
</tr>
<tr>
<td></td>
<td>1.02349</td>
<td>2.76590</td>
<td>35.22837</td>
<td>1.02377</td>
<td>2.5403</td>
<td>33.859</td>
</tr>
<tr>
<td>200</td>
<td>1.01318</td>
<td>2.769357</td>
<td>1.774541</td>
<td>1.01833</td>
<td>2.5866</td>
<td>1.5281</td>
</tr>
<tr>
<td></td>
<td>1.01316</td>
<td>2.73771</td>
<td>1.770475</td>
<td>1.01839</td>
<td>2.5113</td>
<td>1.5338</td>
</tr>
<tr>
<td></td>
<td>1.0132</td>
<td>2.694498</td>
<td>38.7658</td>
<td>1.01183</td>
<td>2.5356</td>
<td>37.878</td>
</tr>
<tr>
<td>300</td>
<td>1.0111</td>
<td>2.78222</td>
<td>1.864912</td>
<td>1.01665</td>
<td>2.6044</td>
<td>1.5739</td>
</tr>
<tr>
<td></td>
<td>1.0108</td>
<td>2.74616</td>
<td>1.866877</td>
<td>1.01683</td>
<td>2.4840</td>
<td>1.5752</td>
</tr>
<tr>
<td></td>
<td>1.00954</td>
<td>2.70656</td>
<td>41.31466</td>
<td>1.01683</td>
<td>2.5149</td>
<td>40.386</td>
</tr>
</tbody>
</table>

4. Numerical Examples

Body Fat dataset is used to evaluate the performance of our proposed method. This dataset contains 20 observations and has three predictors (p=3). Kutner et al. [9] 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.

Table 4: The \( R^2 \) and VIF values for the original Body Fat data set

<table>
<thead>
<tr>
<th>Variables</th>
<th>CVIF</th>
<th>RVIF-MM</th>
<th>FRVIF (GM(DRGP))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R^2 )</td>
<td>( VIF )</td>
<td>( R^2 )</td>
<td>( RVIF )</td>
</tr>
<tr>
<td>( x_1 )</td>
<td>0.9985</td>
<td>708.842</td>
<td>0.84017</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>0.9982</td>
<td>564.343</td>
<td>0.84465</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>0.9904</td>
<td>104.606</td>
<td>0.80516</td>
</tr>
</tbody>
</table>

Table 5: \( R^2 \) and VIF values for modified Body Fat data set

<table>
<thead>
<tr>
<th>Variables</th>
<th>CVIF</th>
<th>RVIF-MM</th>
<th>FRVIF (GM(DRGP))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R^2 )</td>
<td>( VIF )</td>
<td>( R^2 )</td>
<td>( RVIF )</td>
</tr>
<tr>
<td>( x_1 )</td>
<td>0.059848</td>
<td>0.07054</td>
<td>1.06365</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>0.028178</td>
<td>0.02470</td>
<td>1.02899</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>0.051718</td>
<td>0.06607</td>
<td>1.05454</td>
</tr>
</tbody>
</table>

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

5. 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 a fast robust RVIF method for detecting multicollinearity in a data set. The proposed method is formulated by incorporating fast DRGP of Lim and Habshah [14]. 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.

6. References


