An improved robust variance inflation factor: Reducing the negative effects of good leverage points

Osman U. Ekiz
Dept. of Statistics, Gazi University, Turkey
Corresponding author: ufukekiz@gazi.edu.tr

Abstract

In multiple linear regression analysis, the variance inflation factor is a well-known collinearity measure. It is defined as the function of the coefficient of determination between the explanatory variables, and it is based on the maximum likelihood estimator of the regression coefficients. Nevertheless, in addition to outliers, leverage observations can have significant impact on the coefficient of determination, and thereby the variance inflation factor. This study presents an improved robust variance inflation factor estimator that is not affected by these observations. Simulation studies and a real data analysis indicate that the modified robust variance inflation factor estimator performs better than the traditional one.

Keywords: Collinearity-inducing leverage; collinearity-masking leverage; linear regression; outlier; robust statistics

1. Introduction

The multiple linear regression model is used to make inferences about a response variable using explanatory variables, and it is defined as $Y = X\beta + \epsilon$. The maximum likelihood (*ML*) estimator of β, which is known as the best linear unbiased estimator, is expressed as

$$\hat{\beta}_{ML} = \left(X'X\right)^{-1} X'Y,$$

(Graybill, 1961). In the presence of collinearity problem, the well-know ridge regression estimators are proposed (Hoerl & Kennard, 1970). There are many studies in the literature that focus on ridge regression (Dorugade, 2014). Moreover, studies have suggested the use of robust and ridge-type robust estimators if there are outliers, or both collinearity and outliers, in the regression data (Aftab & Chand, 2018; Alshqaq, 2021; Maronna, 2011; Silvapulle, 1991). The presence of both outliers and one or more leverage observations in the data may have an impact on the severity of collinearity. Here, these collinearity-influencing leverage observations are categorized into two groups according to how they affect collinearity. The first group consists of *collinearity-masking leverage* observations. These observations may lead to the misconception that there is no collinearity in the data. For the second group of observations, called *collinearity-inducing leverage* observations, the outcome is just the opposite. They may lead to a misinterpretation of collinearity in the data.

The variance inflation factor ($VIF_{ML} = 1/\left(1 - R^2_{ML}\right)$) is a measure used to make inferences about collinearity. If its value is larger than 10, there is severe collinearity in the data (Gujrati, 2004). R^2_{ML} is the largest coefficient of determination between $X_j, j = 1, ...k$, and the rest of the explanatory variables. If extreme observations are present in the data, these points would impact $\hat{\beta}_{ML}$ and $\bar{y}$, which means the resulting residuals $(y_i - \hat{y})$ might be larger than they are in reality. This leads to the employment of robust determination coefficient to diagnose collinearity by using

$$R^2_r = 1 - \frac{\sum_{i=1}^{n} w_i \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n} w_i \left(y_i - \bar{y}_w\right)^2},$$

where r denotes a robust estimator and $\bar{y}_w = \sum_{i=1}^{n} w_i y_i \big/ \sum_{i=1}^{n} w_i$. The weights, w_i, and predictions, $\hat{y}_i$, are produced by applying a robust regression estimator (Renaud & Victoria-Feser, 2010). However, this estimator performs well in parameter estimations only in the presence of outliers in the *X* or *Y* direction. The calculated value of the robust VIF $\left(VIF_r = 1/(1 - R_r^2)\right)$ based on R_r^2 with *collinearity-inducing leverage* observations, also called good leverage points, leads to the perception that collinearity exists. Note that, here, R_r^2 denotes the largest robust coefficient of determination established by a robust regression estimator between X_j and the remaining explanatory variables. Since *collinearity-inducing leverage* observations have an impact on this estimator, it is important to build an R_r^2 that is strong despite the presence of these points.

This study aims to improve the R_r^2 and VIF_r, which are referred to as the new R_r^2 $\left(newR_r^2\right)$ and new VIF_r $(newVIF_r)$ based on the $newR_r^2$. The severity of collinearity is determined more accurately with the $newVIF_r$, which is also not impacted by *collinearity-inducing leverage* observations. This makes it easier to determine the best estimator for the regression analysis. In Section 2, robust estimators are mentioned to construct new underlined estimators. The suggested approach is introduced in Section 3. The results, using a real data set, are presented in Section 4. Furthermore, this section provides simulation details that allow for comparisons of the estimators utilized. These findings demonstrate that the $newVIF_r$ based on the $newR_r^2$ provides better results compared to the VIF_r. The paper ends with conclusion in Section 5.

2. Robust *LMS*, *LTS*, and *S* estimators

There are various robust estimators for estimating the parameters in multiple linear regression models. In this study, the most common robust estimators the least median of squares $(\hat{\beta}_{LMS})$, least trimmed square $(\hat{\beta}_{LTS})$ (Rousseeuw & Leroy, 1987), and S $(\hat{\beta}_S)$ (Rousseeuw & Yohai, 1984) are employed to determine the performance of the improved estimator $newVIF_r$.

These estimators are calculated from

$$\hat{\beta}_\ell = \left(X' W_{\ell-1} X\right)^{-1} X' W_{\ell-1} Y,$$

where $W_{\ell-1}$ defines the diagonal weight matrix with elements $w\left(r_i\right)$ and the r_i denotes the residuals, $i = 1, ..., n$ (Rousseeuw & Leroy, 1987). Note that for $\hat{\beta}_{LMS}$ and $\hat{\beta}_{LTS}$, $w_i = 1$ when observation $i \in t$th sub-sample. Otherwise, $w_i = 0$. The weights for the S estimator should be established in each iteration by employing Tukey's bi-weight function (Maronna *et al.*, 2006; Rousseeuw & Yohai, 1984).

3. An improved robust *VIF*

The R_r^2 is not affected by the presence of *collinearity-masking leverage* observations. However, it does not yield good results when there are leverage observations that induce collinearity because it is robust only against outliers. In addition, leverage observations that are considered to be good and regular in the direction of $X_{(-j)}$ (the design matrix X excluding the *j*th explanatory variable) can induce collinearity. Thus, a VIF_r that is dependent on R_r^2 would be adversely affected by these observations as well. In order to overcome this negative effect, it is recommended that the *collinearity-inducing leverage* observations be removed from the $X_{(-j)}$ direction before the R_r^2 is calculated. For this purpose, the VIF_r is improved and called the new VIF_r $(newVIF_r)$ (Ekiz, 2021). The detailed description of the algorithm is as follows:

- For each $X_{(-j)}$ compute the robust estimators $\hat{\tau}\left(X_{(-j)}\right)$ and $\hat{\Sigma}_{X_{(-j)}}^{-1}$ of the location and scale parameters, respectively. In this study minimum covariance determinant (*MCD*) estimators are employed (Rousseeuw & Driessen, 1999).

- Compute Mahalanobis distances, MD_i^2 based on $\hat{\tau}\left(X_{(-j)}\right)$ and $\hat{\Sigma}_{X_{(-j)}}^{-1}$ (Maronna *et al.*, 2006).

- If $MD_i^2 > \chi_{k-1,1-\alpha}^2$, x_i is determined to be an *collinearity-inducing leverage* (outlier) observation. Additionally, this point is referred to as good leverage when regressing X_j on $X_{(-j)}$. $\chi_{k-1,1-\alpha}^2$ is the upper-α quantile of the chi-square distribution. At the end of this step, a total of m observations are identified as *collinearity-inducing leverage*.

- Considering that there are *collinearity-inducing leverage* points during the application of the regression of X_j on $X_{(-j)}$, subtract m observations from the data. Both R_r^2 and VIF_r are then computed by constructing the regression analysis with a clean $n-m$ observation.

- Report the estimates from $n-m$ observations as $newR_r^2$ and $newVIF_r$.

When the computed $newVIF_r$ is larger than 10, there is severe collinearity in the data.

4. Application

In this section, the improved measure, $newVIF_r$, is compared with the VIF_r by applying *Body fat* data, (Kutner *et al.*, 2004), which consists of *collinearity-masking leverage* observations. There are three explanatory variables, each of which has 20 observations: Tricep skin thickness (X_1), thigh circumference (X_2), and midarm circumference (X_3).

Let $newVIF_r$ $(r = LMS, LTS, S)$ be the new robust measure, and let VIF_{ML} denote the VIF computed using the *ML* estimator. The values of VIF_r based on LMS, LTS, and S estimators are calculated as 250.2497, 688.5522, and 792.8248, respectively. The values of $newVIF_r$ based on the same estimators are calculated as 825.7449, 790.7602, and 793.9697, respectively. Here, $\alpha = 0.05$. All of these values are much higher than VIF_{ML} which is 36.4631. This is the evidence of the presence of more severe collinearity. Hence, in the case of *collinearity-masking leverage* in the data, the use of VIF_r and $newVIF_r$ estimates will be useful to diagnose the severity of collinearity for the appropriate regression model.

The $newVIF_r$ would not be affected from the *collinearity-inducing leverage* observations existing in the data, in contrast to VIF_r. To illustrate this point of view a detailed simulation study is carried out in Section 4.1. The results both in the application and the simulation study are obtained by using Matlab.

4.1 Simulation study

In this simulation, the datasets are generated so that they are contaminated with leverage observations that effect collinearity. An evaluation of the performance of the VIF_r and $newVIF_r$ estimators with contaminated data is conducted by comparing their Monte Carlo (*MC*) means with the *uVIF* computed from the uncontaminated portion of the data. When the *MC* mean of the estimator is close to the *uVIF*, it can be said that the estimator is not affected by contaminated data (Ekiz, 2021). Note that $uVIF = 1/\left(1 - C_{X_j,X_{(-j)}} C_{X_{(-j)},X_{(-j)}} C'_{X_j,X_{(-j)}}\right)$, where C denotes the correlation matrix of the distribution of the uncontaminated part of the data (Mardia *et al.*, 1979). The datasets are simulated from the contaminated normal distribution, where the number of explanatory variables is set to 3 ($k = 3$). The joint probability distribution of (X_1, X_2, X_3) is defined as $F = (1-\lambda)\,G + \lambda H$, where $G \sim N_k(\mu, \Sigma)$, $H \sim N_k(\theta, \Sigma)$, and $\Sigma = C$. The mixture parameter, $\lambda \in [0, 1]$, provides $\lambda \ll 1$ (Maronna *et al.*, 2006). Additionally, $\mu_X = (\mu_{X_1}, \mu_{X_2}, \mu_{X_3})$ and $\theta_X = (\theta_{X_1}, \theta_{X_2}, \theta_{X_3})$ are used as the location parameters of G and H, respectively. To simulate an n sized dataset consisting of only high-leverage points (masking or inducing) with a proportion of λ, the leverage observations are generated from $N_k(\theta, \Sigma)$ and the non-leverage observations are generated from $N_k(\mu, \Sigma)$. In this way, the set of design parameters $\mu_{X_1}, \mu_{X_2}, \mu_{X_3}, \theta_{X_1}, \theta_{X_2}, \theta_{X_3}$ can be utilized to manipulate the level and type of contamination.

Using a covariance matrix Σ, with ones on the diagonal, the dataset includes *collinearity-masking leverage*. The remaining elements of this matrix are selected as values close to one, providing strong collinearity between the explanatory variables. In the simulations, a λ proportion of high-leverage observations, taken from $H \sim N_k(\theta, \Sigma)$, where $\theta = (5, 7, 7)$, are integrated into the dataset as well. The VIF_{ML_G} and VIF_{ML_F} should be calculated from the observations that are produced from the distributions *G* and *H*, respectively. It can be seen that VIF_{ML_F} is much smaller than VIF_{ML_G}, even for small values of λ. Therefore, a small number of high-leverage observations may mask a strong collinearity that depends on the rest of the data. To create a dataset with *collinearity-inducing leverage*, the elements of Σ are chosen to be very close to zero. Thus, the value of the corresponding *uVIF* is small, indicating that there is no correlation between the explanatory variables. When the λ ratio of the *collinearity-inducing*

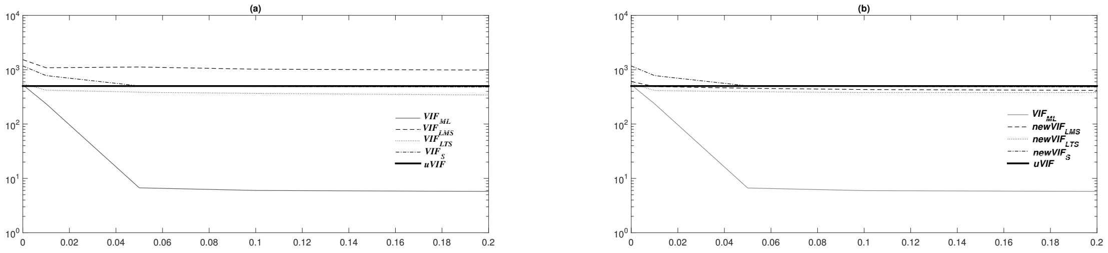

Fig. 1. Contaminated data with *collinearity-masking leverage*. The value of $uVIF$ is set at 501.3193

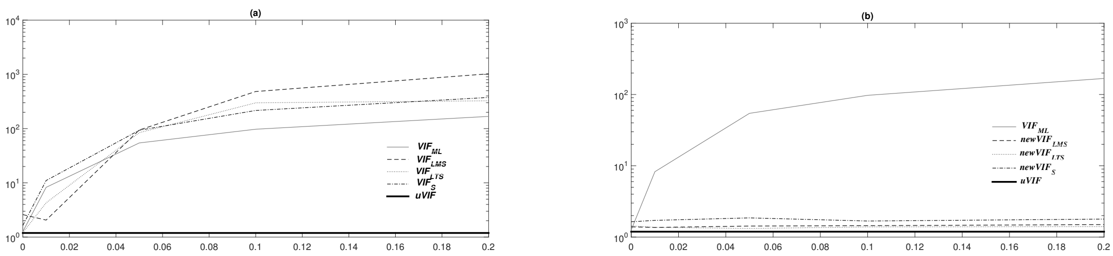

Fig. 2. Contaminated data with *collinearity-inducing leverage*. The value of $uVIF$ is set at 1.2121.

leverage observations generated from the H distribution, with the $\theta = (35, 32, 37)$, is integrated into the data, the calculated VIF_{ML_F} is much higher than the calculated VIF_{ML_G} without the *collinearity-inducing leverage* observations. This result indicates that a small number of *collinearity-inducing leverage* observations may increase the severity of collinearity.

The simulation procedure is based on 10000 iterations for all combinations of $n = 100$ and $\lambda = 0, 0.01, 0.05, 0.10, 0.20$. The *MC* estimations for the VIF_r and $newVIF_r$ values obtained in cases where the data is contaminated by *collinearity-masking* and *-inducing leverage* observations are given in the vertical axes of the graphs in Figure 1 and 2. In these graphs, the horizontal axes show the λ. *MC* estimates near $uVIF = E\,(VIF_{ML_G})$ are considered to be good performance estimates. Note that *E* shows the expected value, and VIF_{ML_G} is the measure of the VIF obtained from the data produced by the *G* distribution, based on the *ML* estimator.

In the case of *collinearity-masking leverage*, the outcomes of both VIF_S and $newVIF_S$ seem to be good (see Figure 1(a) and (b), respectively). Moreover, as shown in Figure 1 and 2, in contrast to the other estimators, the $newVIF_S$ estimator outperforms in both cases, and its calculated values approach $uVIF$.

In the presence of *collinearity-inducing leverage* observations, it can be seen that the VIF_r yields very large results than the $uVIF$. This leads to the misconception of as if there is collinearity, as shown in Figure 2(a). However, according the plots in Figure 2(b) the $newVIF_r$ provides very reasonable results. When $n = 50$, $\lambda = 0.10$, and using the data simulated with *collinearity-inducing leverage* observations, the MC means of VIF_S and $newVIF_S$ are calculated as 350 and 1.80 . Thus, the bias of $newVIF_S$ from $uVIF = 1.2121$ is negligible compared to the value of VIF_S.

5. Conclusion

Before starting a regression analysis, it is important to investigate whether there are outliers and/or collinearity problems in the data. It is recommended that ridge, robust, and ridge-type robust estimators be used for problems with collinearity, outliers, and both collinearity and outliers, respectively (Silvapulle, 1991). Hence, accurately determining the severity of collinearity plays an important role in identifying the correct estimator to apply. When the leverage observations (outliers) in the direction of *j*th explanatory variable mask collinearity (*collinearity-masking leverage*), the results of $V\,IF_r$ demonstrate that there is more severe collinearity in the data, compared to results based on $V\,IF_{ML}$. At the same time, similar results are observed from the proposed *newV IF*$_r$.

However, if the data contains *collinearity-inducing leverage* observations, the VIF_r is unable to recognize that there is actually no collinearity in the data. The VIF_r provides large numerical results, as if collinearity exists. In contrast, the values of the $newVIF_r$ estimator, improved in this study, are small in this situation. Furthermore, when *collinearity-masking* or *-inducing leverage* observations are present in the data, the $newVIF_S$ out-performs the other estimators. For this reason, this measure could be used to diagnose collinearity before deciding which estimator to use for parameter estimates.

References

Aftab, N., & **Chand, S. (2018)**. A simulation-based evidence on the improved performance of a new modified leverage adjusted heteroskedastic consistent covariance matrix estimator in the linear regression model. Kuwait Journal of Science, 45(3).

Alshqaq, S. S. (2021). On the least trimmed squares estimators for *JS* circular regression model. Kuwait Journal of Science, 48(3), 1-13.

Dorugade, A. V. (2014). On comparison of some ridge parameters in ridge regression. Sri Lankan Journal of Applied Statistics, 15(1), 31–45.

Ekiz, O. U. (2021). İyi kaldıraç noktalarından etkilenmeyen sa ̆glam varyans artış faktörü [A variance inflation factor which is robust against leverage points]. II. International Applied Statistics Conference. Proceedings book of the UYIK-2021, 117. Tokat, Turkey.

Graybill, F. A. (1961). Introduction to Linear Statistical Models. New York, USA: McGraw-Hill.

Gujrati, D. N. (2004). Basic Econometrics. New Delhi, IND: Tata McGraw-Hill.

Hoerl, A. E., & **Kennard, R. W. (1970)**. Ridge regression: Biased Estimation for nonorthogonal problems. Technometrics, 12(1), 55–67.

Kutner, M., Nachtsheim, C., & **Neter, J. (2004)**. Applied Linear Regression Models. New York, USA: McGraw-Hill.

Mardia, K. V., Kent, J. T., & **Bibby, J. (1979)**. Multivariate Analysis, New York, USA: Academic Press.

Maronna, R. A., Martin, D. R., & **Yohai, V. J. (2006)**. Robust Statistics: Theory and Methods, New York, USA: Wiley.

Maronna, R. A. (2011). Robust ridge regression for high-dimensional data. Technometrics, 53(1), 44–53.

Renaud, O., & **Victoria-Feser, M. P. (2010)**. A robust coefficient of determination for regression. Journal of Statistical Planning and Inference, 140(7), 1852–1862.

Rousseeuw, P., & **Yohai, V. (1984)**. Robust Regression by Means of S-Estimator. In J. Franke, W. Härdle, & D. Martin (Eds.), Robust and Nonlinear Time Series Analysis (pp. 256–272), New York, USA: Springer.

Rousseeuw, P. J., & **Leroy, A. M. (1987)**. Robust Regression and Outlier Detection. New York, USA: John Wiley & Sons.

Rousseeuw, P. J., & **Driessen, K. V. (1999)**. A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41(3), 212–223.

Silvapulle, M. J. (1991). Robust ridge regression based on an M estimator. Australian Journal of Statistics, 33(3), 319–333.

Submitted: 02/08/2021
Revised: 15/11/2021
Accepted: 16/11/2021
DOI: 10.48129/kjs.15533