Sains Malaysiana 48(9)(2019): 2051–2060
http://dx.doi.org/10.17576/jsm-2019-4809-26

Some New Diagnostics of Multicollinearity in Linear Regression Model

(Beberapa Diagnostik Baru Multikekolinearan dalam Model Regresi Linear)

MUHAMMAD IMDAD ULLAH*, MUHAMMAD ASLAM, SAIMA ALTAF & MUNIR AHMED

ABSTRACT

The problem of multicollinearity compromises the numerical stability of the regression coefficient estimate and cause some serious problem in validation and interpretation of the model. In this paper, we propose two new collinearity diagnostics for the detection of collinearity among regressors, based on coefficient of determination and adjusted coefficient of determination from auxiliary regression of regressors. A Monte Carlo simulation study has been conducted to compare the existing and proposed collinearity diagnostic tests. Comparison of diagnostics on some existing collinear data are also made.

Keywords: Collinearity diagnostics; ill-conditioning; linear dependencies; multicollinearity; regression analysis

ABSTRAK

Masalah multikekolinearan kompromi kestabilan berangka pekali regresi anggaran dan menyebabkan beberapa masalah serius dalam pengesahan dan tafsiran model. Dalam kajian ini, kami mencadangkan dua diagnostik kekolinearan baru untuk pengesanan kekolinearan dalam kalangan peregrasi, berdasarkan pekali penentuan dan pekali penentuan terlaras daripada bantuan regresi oleh peregrasi. Kajian simulasi Monte Carlo telah dijalankan untuk membandingkan kajian kekolinearan sedia ada dengan cadangan ujian kekolinearan diagnostik. Perbandingan diagnostik pada sesetengah data kolinear sedia ada turut dijalankan.

Kata kunci: Analisis regresi; kebergantungan linear; kekolinearan diagnostik; multi-kekolinearan; persuasanaan tak sihat

INTRODUCTION

Consider the usual multiple linear regression model

$$y = X\beta + u,$$

where y is an $n \times 1$ vector of observations on dependent variable; X is known design matrix of order $n \times p$, having full-column rank p; β is a $p \times 1$ vector of unknown parameters and u is an $n \times 1$ vector of random errors with mean zero and variance $\sigma^2 I_n$, where I_n is an identity matrix of order n.

The use and application of the ordinary least squares (OLS) method is popular due to its low computational cost, intuitive plausibility in a wide variety of circumstances and its support by a broad and convoluted body of statistical inference (Belsley et al. 1980). However, linear dependence (relationship; shared variance) between the regressors can affect the model ability to estimate the model's parameters (regression coefficients). Multicollinearity is lack of independence or the presence of interdependence signified by usually high intercorrelations within a set of explanatory variables (Abdullah 1996; Farrar & Glauber 1967; Gunst 1983; Gunst & Månson 1977; Mason et al. 1975). Perfect or near to perfect multicollinearity destroys the uniqueness of the OLS estimators (Belsley et al. 1980).

The OLS estimators can be ambiguous and unstable under severe multicollinearity (i.e. ill-conditioning of $X'X$ matrix). This issue often generates implausible signs, inflated standard errors, low t-ratios with high R-squared (R^2) value, wider confidence intervals, very large condition number and non-significant and/or unexpected magnitude of the regression coefficient estimates. On the basis of theoretical considerations, these indications are thought to be important for detection of multicollinearity among regressors, while the forecasting power of the model may not be affected (Adnan et al. 2006; Belsley et al. 1980; Chen 2012; Greene 2002; Younger 1979).

Many multicollinearity diagnostic indicators are available in the existing literature proposed or discussed by various authors (Belsley 1991; Curto & Pinto 2011; Koutsoyiannis 1978; Kovács et al. 2005; Marquardt 1970; Midi et al. 2011; Montgomery & Askin 1981). Widely used and the most suggested diagnostics are values of pair-wise correlations (Adnan et al. 2006; Chen 2012), variance inflation factor (VIF) and tolerance limit (TOL) (Kutner et al. 2004; Marquardt 1970), eigenvalues values (Kendall 1957; Silvey 1969), condition number (CN) and condition index (CI) (Belsley et al. 1980), Leamer's method (Greene 2002), Klien's rule (Klein 1962), three tests proposed by Farrar and Glauber (1967), Red indicator (Kovács et al. 2005) and Theil's measure (Theil 1971). Table A lists these diagnostics with formulae, references and detection criteria. These collinearity diagnostics are classified and compared as overall (Table 1) and individual (Table 2)

TABLE 1. Percentage detection of collinearity by overall diagnostics measures

n	Indicators	θ				
		0.8366	0.8944	0.9487	0.9747	0.9950
50	Determinant	61.34	99.10	100.00	100.00	100.00
	FGC	100.00	100.00	100.00	100.00	100.00
	Red Indicator	99.90	100.00	100.00	100.00	100.00
	CI	0.00	0.16	44.62	99.32	100.00
	Theil	99.98	100.00	100.00	100.00	100.00
	Sum of reciprocal of eigenvalues	0.46	39.54	99.76	100.00	100.00
100	Determinant	54.34	99.86	100.00	100.00	100.00
	FGC	100.00	100.00	100.00	100.00	100.00
	Red Indicator	100.00	100.00	100.00	100.00	100.00
	CI	0.00	0.00	10.16	100.00	100.00
	Theil	100.00	100.00	100.00	99.82	100.00
	Sum of reciprocal of eigenvalues	0.00	19.48	100.00	100.00	100.00
200	Determinant	48.60	100.00	100.00	100.00	100.00
	FGC	100.00	100.00	100.00	100.00	100.00
	Red Indicator	100.00	100.00	100.00	100.00	100.00
	CI	0.00	0.00	0.28	99.90	100.00
	Theil	100.00	100.00	100.00	100.00	100.00
	Sum of reciprocal of eigenvalues	0.00	5.42	100.00	100.00	100.00

measures of collinearity. The overall diagnostic measures help to get idea about existence of collinearity and result in a single number, while individual measures try to detect the existence of collinearity for each of the regressors.

NEW PROPOSED DIAGNOSTICS

Multicollinearity is considered as a sample phenomenon; therefore, there is no unique method for detection of multicollinearity (Kmenta 1986). So, the existence of multicollinearity should always be tested when examining a data set, in order to avoid the adverse effects of multicollinearity and its pitfall that may exist in regression model. Various diagnostic (graphical and numerical) measures for the quantification of multicollinearity are available in the literature, but none of them can be regarded as a synthetic and normalized indicator at the same time (Curto & Pinto 2011; Green et al. 1978; Kovács et al. 2005; Silvey 1969; Ukoumunne et al. 2002).

In this article, we propose two new diagnostics for multicollinearity. The existing multicollinearity diagnostics depend heavily on R^2 (multiple coefficient of determination) and/or eigenvalues or some relation between R^2 and eigenvalues/ eigenvectors. That is why, the correlation between regressors, the R^2 and eigenvalues are considered as important multicollinearity detection measures.

The proposed collinearity diagnostic measures depend on R^2 and adjusted-R^2 (*adj*-R^2) values from auxiliary regression. The performance of the proposed measures has been evaluated through empirical results using the Monte Carlo simulations. These simulations have been carried out for both uncorrelated and correlated regressors at different levels of correlations and different sample sizes. Some threshold values for the new proposed diagnostics have also been determined.

The R^2 indicates that how well data fit a statistical model as it is the proportion of explained variation in dependent variable due to independent variables. The higher the R^2 value, the more chances of regressors to be plagued with multicollinearity (Asteriou & Hall 2007; Gujarati & Porter 2008; Maddala 1988). The R^2 is a monotone non-decreasing function of number of regressors included in the model. It means R^2 inflates the estimate of how well the regression fits the data (Gujarati & Porter 2008; Stock & Watson 2010). The *adj*-R^2 is a modified version of R^2 (due to Theil 1961) that adjusts for number of regressors in a model relative to the number of data points and hence, it is an attempt to take account of the phenomenon of the automatically and spuriously increasing when extra regressors are added to the model (Stock & Watson 2010). In other words, it deflates the by some factor, i.e., $\frac{n-1}{n-p-1}$. For $p > 1$, *adj*-$R^2 \leq R^2$, implies that as the number of regressor(s) increases, the *adj*-R^2 increases less than the (un-adjusted) R^2, because R^2 is affected by regressors sharing their variances, since linear dependence exists among regressors (Gujarati & Porter 2008; Maddala 1988). The above discussion about R^2 and *adj*-R^2 is the main reason to consider *adj*-R^2 in new diagnostic measures.

In the auxiliary regression, for every regressor the association is checked with the other (remaining) regressors of the model. For this paper, we generated six correlated regressors with various combination of sample size and degrees of correlation among these regressors. For our proposed diagnostic measures, six auxiliary regression models are carried out and coefficient of determination (R_j^2) and adjusted coefficient of determination (adj-R_j^2) are obtained from each regression. The reason of using R_j^2 and adj-R_j^2 is discussed above. The other reason of using R_j^2 and adj-R_j^2 is that stronger the undesired association between the regressor say X_1 with the remaining regressors of the model, more the chances of multicollinearity exists when two or more regressors correlated. Therefore, the degree of multicollinearity can also be expressed by R_j^2 (obtained from auxiliary regression of the *j*th regressor as dependent variable), since the VIF is also built on the idea of auxiliary regression. If the R_j^2 value of any regressor is close to zero, the VIF will be closer to one; hence, no multicollinearity exists in this case. On the other hand, if R_j^2 from an auxiliary regression is very large the VIF would also be large showing severe multicollinearity (Cleff 2013).

From empirical results of the Monte Carlo experiment, existing theory related to coefficient of determinations, inflation (spurious increase) in R^2 values due to addition of regressor(s) in model, and deflation in *adj-*R^2 by factor $\frac{n-1}{n-p-1}$, we suggest to take difference of *adj-*R^2 and from auxiliary regression of regressors to account the sharing of variances due to different regressors in each auxiliary regression run, for the detection of multicollinearity (see Asteriou & Hall 2007; Gujarati & Porter 2008; Maddala 1988, for auxiliary regression). The difference of R_j^2 and *adj-*R_j^2 is used as a new diagnostic measure and is referred to as Indicator 1 (IND1_j) for further discussion.

$$\text{IND1}_j = R_j^2 - adj\text{-}R_j^2 = \frac{(n-1)(1-R_j^2)}{n-p} + R_j^2 - 1,$$

$$= (R_j^2 - 1) \times \left(\frac{1-p}{n-p}\right), \qquad (1)$$

where R_j^2 and *adj-*R_j^2 are from the auxiliary regression of each explanatory variables.

For simulated collinear and non-collinear data, using auxiliary regression, we empirically found that smaller the difference or alternatively closer the value of R_j^2 and *adj-*R_j^2 (R_j^2 – *adj-*$R_j^2 \leq 0.020$), greater the chances of multicollinearity. Alternatively, larger the value of $(R_j^2 - adj\text{-}R_j^2)^{-1} \geq 50$ more severe the multicollinearity will be there. This difference of R_j^2 and *adj-*R_j^2 from auxiliary regression of explanatory variables lies in an interval [0.0104, 0.0418] for various combination of sample size and correlation level between generated regressors. Any of the extreme difference value from the interval can be used as criterion but we used central value (average of value of differences for all sample sizes and correlation levels) which was approximately 0.020.

From (1), as $n \to \infty$, IND1_j approaches to 0. Therefore, multicollinearity is detected when

$$\begin{cases} IND1_j < C & \text{for } n < 100, \\ IND1_j < \dfrac{C}{n} \times 100 & \text{for } n > 100, \end{cases}$$

where, $C \in [0.01, 0.04]$.

The second diagnostic tool is the ratio of each R^2 from the auxiliary regression (that is, R_j^2) to the mean of all R_j^2 i.e., $\frac{R_j^2}{m}$ where $m = \frac{\sum_{j=1}^{p} R_j^2}{p}$, and $j = 1, 2, \ldots, p$. If this ratio for *j*th variable is greater than R^2 (from regression of *y* on *X*'s) then the *j*th regressor will be highly collinear with others regressors. In denominator of this diagnostic, mean of all R_j^2 (*m*) gives the average sharing of variances among regressors accounted by using auxiliary regression for *j*th regressor as dependent variable on the remaining regressors, whereas the distribution of R_j^2 for different sample size and correlation level between variables was found to be approximately normally distributed. Note that if correlation among regressors is small then this proposed indicator (say IND2_j for further reference) will give false positive (wrong) detection of collinearity, as magnitude of $\frac{R_j^2}{m}$ will be larger than the average of R_j^2's $(j = 1, 2, \ldots, p)$ in this case. Since the classic symptom of multicollinearity is $R^2 \geq 0.7$, therefore, to avoid the false positive detection of multicollinearity, the IND2_j specifies multicollinearity when,

$$\text{IND2}_j = \begin{cases} \dfrac{|R_j^2 - 1|}{m} > R^2, & \text{if } 0.7 \leq R^2 < 0.80 \\ \dfrac{R_j^2}{m} > R^2, & \text{if } R^2 \geq 0.80 \\ \textit{no collinearity}, & \text{if } R^2 < 0.70 \end{cases}$$

Thus, the chief objective of this paper was to compare the existing and proposed multicollinearity diagnostic tools for their performance of detection under various combination of level of correlation and sample size.

NUMERICAL EVALUATION

For the numerical evaluation of different diagnostic measures of multicollinearity, we have followed the similar Monte Carlo schemes as used by many other researchers (Aslam 2014; Clark & Troskie 2006; Månsson et al. 2010; McDonald & Galarneau 1975; Newhouse & Oman 1971).

The simulation deals with six parameter case. The explanatory variables are computed as

$$x_{ij} = (1-\theta^2)^{½}\ z_{ij} + \theta\, z_{i7};\ i = 1,2,\ldots,n;\ j = 1,2,\ldots,7,$$

where $z_{i1}, z_{i2}, \ldots, z_{i7}$ are independent standard normal pseudorandom numbers, and correlation between any explanatory variable is given by θ^2. Without loss of generality, these variables are standardized so that *X'X* from a usual correlation matrix. Five different sets of correlations are considered corresponding to $\theta = 0.8366$, 0.9844, 0.9487, 0.9747 and 0.9950. The values of such generated predictors are kept fixed for simulation.

The sample size (*n*) is set to 50, 100, and 200. The number of Monte Carlo replications is set to be 5,000. In addition to simulation study, for illustration purpose, different diagnostic measures were also evaluated on some popular collinear datasets, available in few previous studies (Hald 1952; Longley 1967; Malinvaud 1968). All the computations are performed making programming routines (available as R package mctest (Imdad & Aslam 2018)).

Table 1 contains the simulated results for the overall measure of collinearity diagnostics in percentage of detection that indicates the collinearity among all the regressors. It can be seen that the determinant of $X'X$, the Farrar-Glauber Chi-square (FGC) test, red indicator and Theil's measure detect collinearity correctly than the CI and sum of reciprocal of eigenvalues for all $\theta \geq 0.8944$ and for different sample size ($n = 50$, 100, and 200) while only determinant detects the collinearity poorly for $\theta = 0.8366$. Percentage of detection by the CI is the lowest than all the other overall diagnostics for different sample sizes, but it detects well as θ increases than $\theta = 0.9487$ while for $\theta \geq 0.9747$ detection becomes 100% for all sample sizes. For $\theta = 0.8366$ and $\theta = 0.8944$ the sum of reciprocal of eigenvalues diagnostic detects existence of collinearity among regressors at low percentage, but relatively much higher than that by the CI. The FGC and Theil's indicator successfully diagnose the collinearity between the explanatory variables.

Table 2 consists of the simulated results for collinearity diagnostics for each regressor x_j, referred to as individual measure of diagnostics in the available literature. For $\theta \geq 0.9487$ and sample size $n = 50$, 100 and 200, all the diagnostic measures successfully detect the collinearity among regressors x_j, except VIF/TOL, Leamer's measure and CVIF (Curto & Pinto 2011). For correlation level $\theta = 0.8366$, and 0.8944, the diagnostic measures VIF (or alternatively TOL), CVIF and Leamer's method could not successfully detect the collinearity among regressors. For sample size of 50, the percentages of detection by VIF/ TOL and Leamer's method (when $\theta = 0.8366$) is less than approximately 4% and 17%, respectively. For $n = 50$ and $\theta = 0.8944$ percentage detection by VIF/ TOL and Leamer's method is less than 40% and 74%, respectively. For sample of size 100, the percentages of detection by VIF/ TOL and Leamer's method (when $\theta = 0.8366$) is less than 1% and 4.2%, respectively. Similarly, for $\theta = 0.8944$, the percentage detection is less than 25% and 71%, respectively. Percentage of collinearity detection by CVIF indicator is smaller as compared to the other indicators, as this percentage for sample of size 50 and $\theta = 0.8366$ is less than 1% for $\theta = 0.8944$ is less than 3% and for $\theta = 0.9487$ is less than 41%. The percentage of detection by CVIF indicator increases as correlation among regressors and the sample size both increases. It is worthy to note that the percentage of detection decreases with the increase of sample size which follows the theory that collinearity reduces with the increase of sample size.

On the other hand, our proposed collinearity diagnostics ($IND1_j$ and $IND2_j$) detect 100% existence of collinearity between regressors x_j for different samples size and correlation levels. When the regressors are collinear at $\theta = 0.8366$ and sample size of 50, 100 and 200, the percentage of collinearity detection is less than 65%, 75% and 84%, respectively, by $IND1_j$, while $IND2_j$ detects 100% existence of collinearity for different correlation level and sample sizes. For $\theta \geq 0.8944$, the percentage of detection is about 100%. Thus, when collinearity is needed to be detected rightly, the new proposed measures do it correctly.

We also performed simulation on very large sample size (n = 500, 1000, 2000) with very high or low correlation level ($\theta = 0.3162$, 0.5477, 0.7071, and 0.9999) among regressors. For $n = 100$ and $\theta = 0.5477$, among the overall diagnostic tools, Theil's measure and FGC result in 100% false positive collinearity detection. Among the individual diagnostic measures, the Farrar w_i, F-test, and Klein's rule detected collinearity in most of the cases, reflecting very high false positive rate. On the other hand, the new proposed indicators, $IND1_j$ and $IND2_j$ also detect collinearity about 10% of the times. These results are not presented due to huge volume of diagnostics output.

In Table 3, we tested all collinearity diagnostics on already existing and tested data available in literature. The results indicate that whether different collinearity diagnostic tools detected the collinearity or they failed to detect the collinearity among regressors for three different existing datasets already available in literature. The datasets by Hald (1952), Longley (1967), and Malinvaud (1968), extremely plagued with multicollinearity, were used. All of the overall diagnostic measures successfully detected the existence of collinearity among regressors for these datasets except Theil's measure for Malinvaud data set. Individual diagnostic measures, Klein's rule and CVIF failed to detect the collinearity among regressors for the Longley and Hald datasets. However, Farrar and Glauber's w_i and F-test also detected the existence of collinearity due regressor x_5 for Longley dataset which was not reported by other indicators. New proposed indicators ($IND1_j$ and $IND2_j$) correctly detected the existence of collinearity among regressors x_5 for all three existing datasets. Correct detection by these new indicators also followed the results from the existing literature (Chatterjee & Hadi 2006; Gujarati & Porter 2008; Maddala 1988).

Conclusion

The simulated results favour the use of determinant of normalized correlation matrix without intercept and Red indicator as overall detection of collinearity among regressors, while CN or CI, FGC test, Theil's measure and sum of reciprocal of eigenvalues may be avoided due to their poor detecting behavior. The VIF/TOL and Leamer's method may be used especially if interdependence among regressors is ≥ 0.8944. However, Farrar and Glauber's w_i test, F-test, Klein's rule and CVIF may not be preferred because Farrar and Glauber's tests are criticized by many researchers (Haitovsky 1969; Kumar 1975; O'Hangan & McCabe 1975) and because of high false positive detection by the other diagnostic measures.

TABLE 2. Percentage detection of collinearity by individual diagnostics (for each regressor)

Indicators	**θ = 0.8366**						**θ = 0.8944**						**θ = 0.9487**						**θ = 0.9747**						**θ = 0.9950**					
	X_1	X_2	X_3	X_4	X_5	X_6	X_1	X_2	X_3	X_4	X_5	X_6	X_1	X_2	X_3	X_4	X_5	X_6	X_1	X_2	X_3	X_4	X_5	X_6	X_1	X_2	X_3	X_4	X_5	X_6
n = 50																														
VIF	3.20	2.98	3.50	3.40	3.34	3.54	38.58	38.86	38.26	39.96	38.94	39.00	98.94	99.08	98.94	98.86	99.06	98.96	100	100	100	100	100	100	100	100	100	100	100	100
TOL	3.20	2.98	3.50	3.40	3.34	3.54	38.58	38.86	38.26	39.96	38.94	39.00	98.94	99.08	98.94	98.86	99.06	98.96	100	100	100	100	100	100	100	100	100	100	100	100
Farrar	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Leamer	15.72	15.56	16.76	16.32	16.12	16.10	71.32	73.40	71.24	72.54	71.94	71.30	99.96	99.92	99.94	99.92	99.98	99.94	100	100	100	100	100	100	100	100	100	100	100	100
F-test	100	99.98	99.98	99.98	100	99.98	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Klein	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
CVIF	0.44	0.44	0.50	0.42	0.44	0.46	2.30	2.14	2.06	2.26	2.34	2.32	40.66	40.10	39.60	40.50	40.52	40.32	99.70	99.70	99.70	99.70	99.70	99.70	99.62	99.62	99.62	99.62	99.62	99.62
IND1	**64.56**	**64.82**	**65.62**	**65.28**	**64.80**	**64.84**	**97.68**	**97.46**	**97.58**	**97.54**	**97.70**	**97.84**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
IND2	**99.84**	**99.82**	**99.78**	**99.88**	**99.80**	**99.92**	**99.86**	**99.96**	**99.90**	**99.82**	**99.92**	**99.84**	**99.94**	**99.98**	**99.98**	**99.90**	**99.90**	**99.96**	**99.98**	**100**	**100**	**99.98**	**100**	**99.98**	**99.96**	**99.98**	**99.90**	**99.90**	**99.94**	**99.90**
n = 100																														
VIF	0.16	0.08	0.10	0.28	0.18	0.10	23.86	24.92	24.04	24.52	24.04	24.58	99.90	99.78	99.88	99.88	99.94	99.90	100	100	100	100	100	100	100	100	100	100	100	100
TOL	0.16	0.08	0.10	0.28	0.18	0.10	23.86	24.92	24.04	24.52	24.04	24.58	99.90	99.78	99.88	99.88	99.94	99.90	100	100	100	100	100	100	100	100	100	100	100	100
Farrar	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Leamer	3.90	3.42	4.10	3.96	4.18	3.92	70.84	71.40	70.56	70.38	70.54	70.46	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
F-test	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Klein	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
CVIF	0	0	0	0	0	0	0.06	0.06	0.04	0.06	0.06	0.06	26.96	27.18	26.50	26.68	26.74	26.68	99.82	99.70	99.62	99.56	99.76	99.74	**99.98**	**99.98**	**99.98**	**99.98**	**99.98**	**99.98**
IND1	**73.72**	**73.92**	**74.24**	**72.74**	**73.80**	**74.26**	**99.84**	**99.82**	**99.82**	**99.84**	**99.78**	**99.90**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
IND2	**100**	**100**	**100**	**100**	**99.98**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
n = 200																														
VIF	0	0	0	0	0	0	11.92	11.62	12.12	11.52	12.32	12.16	100	100	99.98	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
TOL	0	0	0	0	0	0	11.92	11.62	12.12	11.52	12.32	12.16	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Farrar	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Leamer	0.54	0.22	0.26	0.20	0.44	0.38	73.32	72.44	72.58	72.70	71.84	72.60	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
F-test	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Klein	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
CVIF	0	0	0	0	0	0	0.08	0.12	0.16	0.12	0.12	0.12	13.16	14.02	13.34	13.45	13.24	14.36	99.98	99.96	99.98	100	99.98	100	100	100	100	100	100	100
IND1	**83.00**	**83.24**	**84.06**	**83.12**	**83.56**	**83.38**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
IND2	**100**	**100**	**100**	**100**	**99.98**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

TABLE 3. Collinearity detection by overall and individual indicators for existing collinear datasets

Diagnostic	Data Set	Indicators	Results *						
Overall	Longley	Determinant	1						
		Farrar	1						
		Red Indicator	1						
		CI	1						
		Theil	1						
		Sum of reciprocal of eigenvalues	1						
Individual				X1	X2	X3	X4	X5	X6
		VIF		1	1	1	1	0	1
		TOL		1	1	1	1	0	1
		Farrar		1	1	1	1	1	1
		Leamer		1	1	1	1	0	1
		F-test		1	1	1	1	1	1
		Klein		1	0	1	0	0	1
		CVIF		0	0	0	0	0	0
		IND1		1	1	1	1	0	1
		IND2		**1**	**1**	**1**	**1**	**0**	**1**
Overall	Malinvaud	Determinant	1						
		Farrar	1						
		Red Indicator	1						
		CI	1						
		Theil	0						
		Sum of reciprocal of eigenvalues	1						
Individual				X1	X2	X3			
		VIF		1	0	1			
		TOL		1	0	1			
		Farrar		1	0	1			
		Leamer		1	0	1			
		F-test		1	0	1			
		Klein		1	0	1			
		CVIF		1	0	1			
		IND1		1	0	1			
		IND2		1	0	1			
Overall	Hald	Determinant	1						
		Farrar	1						
		Red Indicator	1						
		CI	1						
		Theil	1						
		Sum of reciprocal of eigenvalues	1						
Individual				X1	X2	X3	X4		
		VIF		1	1	1	1		
		TOL		1	1	1	1		
		Farrar		1	1	1	1		
		Leamer		1	1	1	1		
		F-test		1	1	1	1		
		Klein		0	1	0	1		
		CVIF		0	0	0	0		
		IND1		1	1	1	1		
		IND2		1	1	1	1		

* 1 indicates that collinearity is detected by the indicator while 0 indicates the failure of detection

Among the individual diagnostic measures, Klein's rule and our proposed indicators ($IND1_j$ and $IND2_j$) are recommended for detection of collinearity. The measures, $IND1_j$ and $IND2_j$ are reported to give attractive performance for successful detection of linear dependencies among regressors for different level of correlation among regressors and samples sizes.

Our proposed collinearity diagnostic $IND1_j$ should be preferred over $IND2_j$ and the other diagnostics as it correctly detects the collinearity among regressors at different sample sizes and correlation level among regressors. $IND2_j$ may gave false positive detection for small samples and low correlation among regressors.

REFERENCES

Abdullah, M.B. 1996. Detection of influential observations in principle component regression. *Sains Malaysiana* 25(1): 145-160.

Adnan, N.A., Maizah, H. & Adnan, R. 2006. A comparative study on some methods for handling multicollinearity problems. *Mathematika* 22(2): 109-119.

Aslam, M. 2014. Using heteroscedasticity-consistent standard errors for the linear regression model with correlated regressors. *Communications in Statistics-Simulation and Computation* 43(10): 2353-2373.

Asteriou, D. & Hall, S.G. 2007. *Applied Econometrics: A Modern Approach using Eviews and Microfit*. New York: Palgrave Macmillan.

Belsley, D.A. 1991. A guide to using the collinearity diagnostics. *Computer Science in Economics and Management* 4(1): 33-50.

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

Chatterjee, S. & Hadi, A. 2006. *Regression Analysis by Example*. 4th ed. New York: John Willey & Sons Inc.

Chen, G.J. 2012. A simple way to deal with multicollinearity. *Journal of Applied Statistics* 39(9): 1893-1909.

Clark, A.E. & Troskie, C.G. 2006. Ridge regression: A simulation study. *Communications in Statistics - Simulation and Computation* 35(3): 605-619.

Curto, J.D. & Pinto, J.C. 2011. The Corrected VIF (CVIF). *Journal of Applied Statistics* 38(7): 1499-1507.

Dillon, W.R. & Goldstein, M. 1984. *Multivariate Analysis: Methods and Applications*. New York: John Wiley & Sons, Inc.

Farrar, D.E. & Glauber, R.R. 1967. Multicollinearity in regression analysis: The problem revisited. *The Review of Economics and Statistics* 49(1): 92-107.

Green, P.E., Carrol, J.D. & DeSarbo, W.S. 1978. A new measure of predictor variable importance in multiple regression. *Journal of Marketing Research* 15(3): 356-360.

Greene, W.H. 2002. *Econometric Analysis.* 5th ed. New Jersey: Prentice Hall.

Gujarati, D.N. & Porter, D.C. 2008. *Basic Econometrics*. 5th ed. New York: McGraw-Hill.

Gunst, R.F. 1983. Regression analysis with multicollinear predictor variables: Definition, detection and effects. *Communications in Statistics - Theory and Methods* 12(19): 2217-2260.

Gunst, R.F. & Måson, R.L. 1977. Advantages of examining multicollinearities in regression analysis. *Biometrics* 33(1): 249-260.

Haitovsky, Y. 1969. Multicollinearity in regression analysis: Comment. *The Review of Economics and Statistics* 51(4): 486-489.

Hald, A. 1952. *Statistical Theory with Engineering Applications*. New York: John Wiley & Sons.

Imdad, M.U. & Aslam, M. 2018. *mctest: Multicollinearity Diagnostic Measures*. https://CRAN.R-project.org/package=mctest, version 1.2.

Kendall, M.G. 1957. *A Course in Multivariate Analysis*. Griffin: London. pp. 70-75.

Klein, L.R. 1962. *An Introduction to Econometrics*. 2nd ed. New Jersey: Prentice-Hall.

Kmenta, J. 1980. *Elements of Econometrics*. Macmillan Publishing Company: New York.

Koutsoyiannis, A. 1978. *Theory of Econometrics*. 2nd ed. Maryland: Rowman & Littlefield Publishers.

Kovács, P., Petres, T. & Tóth, L. 2005. A new measure of multicollinearity in linear regression models. *International Statistical Review/Revue Internationale de Statistique* 73(3): 405-412.

Kumar, K.T. 1975. Multicollinearity in regression analysis. *The Review of Economics and Statistics* 57(3): 365-366.

Kutner, M.H., Nachtsheim, C.J. & Neter, J. 2004. *Applied Linear Regression Models*. 4th ed. New York: McGraw-Hill.

Longley, J.W. 1967. An appraisal of least squares programs for the electronic computer from the viewpoint of the user. *Journal of the American Statistical Association* 62: 819-841.

Maddala, G.S. 1988. *Introduction to Econometrics*. New York: Macmillan.

Malinvaud, E. 1968. *Statistical Methods of Econometrics*. Amsterdam: North-Holland Publication. pp. 187-192.

Månsson, K., Shukur, G. & Kibria, B.M.G. 2010. A simulation study of some ridge regression estimators under different distributional assumptions. *Communications in Statistics-Simulation and Computation* 39(8): 1639-1670.

Marquardt, D.W. 1970. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics* 12(3): 591-612.

Mason, R.L., Gunst, R.F. & Webster, J. 1975. Regression analysis and problems of multicollinearity. *Communications in Statistics* 4(3): 277-292.

McDonald, G.C. & Galarneau, D.I. 1975. A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association* 70(350): 407-416.

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

Montgomery, D.C. & Askin, R.G. 1981. Problems of nonnormality and multicollinearity for forecasting methods based on least squares. *AIIE Transactions (American Institute of Industrial Engineers)* 13(2): 102-115.

Newhouse, J.P. & Oman, S.D. 1971. *An Evaluation of Ridge Estimator*. Rand Report, No. R-176-PR.

O'Hagan, J. & McCabe, B. 1975. T-Tests for the severity of multicollinearity in regression analysis: A comment. *The Review of Economics and Statistics* 57(3): 368-370.

Silvey, S.D. 1969. Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society. Series B* 31(3): 539-552.

Stock, J.H. & Watson, M.W. 2010. *Introduction to Econometrics*. 3rd ed. Boston: Pearson Addison-Wesley.
Theil, H. 1971. *Principles of Econometrics*. New York: John Wiley & Sons.
Ukoumunne, O.C., Gulliford, M.C. & Chinn, S. 2002. A note on the use of the variance inflation factor for determining sample size in cluster randomized trials. *Journal of the Royal Statistical Society. Series D (The Statistician)* 51(4): 479-484.
Younger, M.S. 1979. *A Handbook for Linear Regression*. North Scituate, MA: Duxbury Resource Center.

Muhammad Imdad Ullah*, Muhammad Aslam & Saima Altaf
Department of Statistics
Bahauddin Zakariya University
Multan 60800
Pakistan

Munir Ahmed
Department of Management Sciences
COMSAT University, Vehari Campus
Islamabad

*Corresponding author; email: mimdadasad@gmail.com

Received: 21 October 2018
Accepted: 3 May 2019

APPENDIX

TABLE A. Listing of collinearity diagnostics

Diagnostic	Description, formula and cutoff	Criteria	References
Correlation Matrix	High zero order or Pairwise correlation between regressors	$r_{ij} > 0.8$	Adnan et al. 2006; Gujarati & Porter 2008; Maddala 1988
Determinant	Determinant of normalized correlation matrix without intercept, while $0 \leq \lvert X'X \rvert \leq 1$	$\lvert X'X \rvert \sim 0$	Asteriou & Hall 2007
R^2, Var(β)'s & t-ratios	High R^2 value, conversely high variance of β's and low t-ratios		Gujarati & Porter 2008; Maddala 1988
Farrar χ^2	$\chi^2 = -\left[n-1-\frac{1}{6(2p+5)}\right] \times \log_e\left[X'X\right] \sim \psi^2_{v=\frac{1}{2}p(p-1)}$	$\chi^2 > \psi^2_{\frac{1}{2}p(p-1)}$	Farrar & Glauber 1967
Farrar w_i	$w_i = \frac{R_j^2}{1-R_j^2}\left(\frac{n-p}{p-1}\right) \sim F(n-p, p-1)$	$w_i > F_{(n-p,\, p-1)}$	Farrar & Glauber 1967
Klein's Rule	If $R^2_{xj.x_1,x_2,\ldots,x_p} > R^2_{y.x_1,x_2,\ldots,x_p}$, multicollinearity may be troublesome.		Klein 1962
VIF and TOL	$\left(X'X\right)^{-1}_{jj} = VIF_j = \frac{1}{1-R_j^2}$ $TOL_j = \frac{1}{VIF_j} = 1 - R_j^2$	VIF > 3, 5, 10 TOL ~ 0	Kutner et al. 2004; Marquardt 1970
Eigenvalues	Smaller eigenvalues of $X'X$ or its related correlation matrix indicate collinearity.	Relatively smaller than other eigenvalues	Kendall 1957; Silvey 1969
CI	$CI_j = \sqrt{\frac{\max(\lambda_j)}{\lambda_j}}; j = 1,2, \ldots, p;\ \lambda_1 \geq \lambda_2 \ldots \lambda_p$	$CI_j > 10, 15, 30$	Belsley 1980; Chatterjee & Hadi 2006; Maddala 1988
Sum of λ_j^{-1}	$\sum_{j=1}^{p} \frac{1}{\lambda_j}; j = 1,2,\ldots,p$	five times the number of predictors	Chatterjee & Hadi 2006; Dillon & Goldstein 1984
CVIF	$CVIF_j = VIF_j \times \frac{1-R^2}{1-R_0^2}$; $R_0^2 = R^2_{yx_1} + R^2_{yx_2} + \ldots + R^2_{yx_p}$	$CVIF_j \geq 10$	Curto & Pinto 2011
Leamer	$C_j = \left\{\frac{\left(\sum_i^n \left(X_{ij} - \bar{X}_j\right)^2\right)^{-1}}{\left(X'X\right)^{-1}_{jj}}\right\}^{\left(\frac{1}{2}\right)}$	$C_j \sim 0$	Greene 2002

Continue TABLE A.

Diagnostic	Description, formula and cutoff	Criteria	References
Theil's indicator	$m = R^2 - \sum_{j=1}^{p}\left(R^2 - R_{-i}^2\right)$	$m \sim 1$ $m \sim 0$, no redundancy	Theil 1971
Red indicator	$Red = \dfrac{\sqrt{\sum_{j=1}^{p}(\lambda_j - 1)^2}}{\dfrac{p}{\sqrt{p-1}}}$	$Red \sim 1$, collinearity	Kovács et al. 2005
F and R^2 Relation	$F_i = \dfrac{\dfrac{R^2_{xj.x_1,\dots,x_p}}{p-2}}{\dfrac{1-R^2_{xj.x_1,\dots,x_p}}{n-p+1}} \sim F(p-2, n-p+1)$	$F_i > F^*$ $F^* = F_{p-2,\, n-p+1}$	Gujarati & Porter 2008
IND1_j	$IND1_j = R_j^2 - \text{adj-}R_j^2$	$\begin{cases} IND1_j < C & \text{for } n < 100 \\ IND1_j < \dfrac{C}{n} \times 100 & \text{for } n > 100 \end{cases}$ where $C \in [0.01, 0.04]$	(present article)
IND2	$IND2_j = \dfrac{R_j^2}{m};\ j = 1, 2, \dots, p;\ m = \dfrac{\sum_j R_j^2}{p}$	$\begin{cases} \dfrac{\lvert R_j^2 - 1\rvert}{m} > R^2, & \text{if } 0.70 \le R^2 < 0.80, \\ \dfrac{R_j^2}{m} > R^2, & \text{if } R^2 \ge 0.80, \\ \text{no collinearity}, & \text{if } R^2 < 0.70 \end{cases}$	(present article)