國立臺南大學
「教育研究學報」第 57 卷第 2 期（民國 112 年）：67~92
DOI：10.53106/199044282023105702004

Why a Variance Inflation Factor of 10 Is Not an Ideal Cutoff for Multicollinearity Diagnostics

Cheng-Chang Jeng*

ABSTRACT

In linear regression analysis, the variance inflation factor (*VIF*) is often used to determine whether multicollinearity exists among independent variables (IVs). Despite its frequent use, no consensus has been achieved regarding a *VIF* threshold that reliably indicates multicollinearity. Although researchers have historically indicated that cutoff values ranging from 2 to 10 should be used, no single value has gained universal acceptance. To address this problem, this study used an R-based platform to calculate *VIF* values under various conditions, including different numbers of IVs (denoted as *k*) and paired correlation coefficients between IVs (denoted as *r*). The study discovered that *VIF* values are influenced by both the number of IVs and the degree of correlation between them. Moreover, when the correlation coefficient is held constant and the number of IVs increases infinitely, the *VIF* tends to converge at a limit. The study also asserts that employing a universal *VIF* cutoff for multicollinearity detection is impractical because the cutoff must be determined with consideration of both the specific number of IVs in a linear regression model and the correlation coefficients researchers deem to be acceptable. The study developed a table of *VIF* cutoff values to aid researchers in identifying suitable cutoff values for their linear regression analyses. The study concludes by discussing its limitations.

Keywords: Linear Regression, Multicollinearity, Variance Inflation Factor (VIF)

*Department of Education, National Taitung University
Corresponding Author: Cheng-Chang Jeng, e-mail: jengcc@gmail.com

I. Introduction

This section first discusses the literature review of linear regression analysis and defines the variance inflation factor (*VIF*) and other indexes commonly used in multicollinearity diagnostics. Various criteria and issues associated with using *VIF* are also explored, and then the motives and objectives of this study are presented.

A. Literature Review

(A) Defining R^2, *Tolerance*, and *VIF*

The term "multiple linear regression analysis" is conveniently referred to as "linear regression analysis" in this study. A linear regression analysis is a crucial form of statistical analysis in several fields. Its purpose can be roughly divided into three categories (Chiou, 2021; Lin, 2014; Yen, 1994): (1) establishing prediction models, (2) exploring the strength and direction of associations between independent variables (IVs) and a dependent variables (DV), and (3) observing the trends in time series. Due to its wide range of applications, linear regression analyses are often utilized in quantitative research in the field of education.

In a linear regression analysis, reducing the correlations among IVs enables researchers to interpret the prediction model with precision. Moderate or high degrees of correlation among IVs represent multicollinearity. The primary dangers of multicollinearity are as follows (Lewis-Beck, 1980): (1) a regression equation may have a fairly high R^2 value, but the coefficients of the IVs are not significant; (2) the coefficients of some IVs may change drastically with adding or removing other IVs; (3) the coefficients of the IVs will become unstable, with extremely high (or low) values which should indicate significance (or insignificance), but these indicators are not reliable, making the model difficult to explain; (4) interpretations of the polarity (positive or negative sign) of the IV coefficients may be the opposite of the norm. Diagnosing whether multicollinearity exists among the IVs is thus a crucial issue in a linear regression analysis.

Tools commonly used to detect multicollinearity include Pearson's correlation coefficient, R^2, *tolerance*, and *VIF*. The latter three are defined as follows (Darmawan & Keeves, 2006; Hair et al., 2006; James et al., 2013):

A regression equation with n observed values and *j* IVs is as shown in Equation (1):

$$\hat{y} = a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_jx_j + c. \quad (1)$$

Thus, the coefficient of determination of $\hat{y}$, also known as explanatory power R^2, is defined as shown in Equation (2):

$$R^2 = \frac{SS_{regression}}{SS_{total}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \quad (2)$$

Because Equation (3) is established,

$$SS_{total} = SS_{regression} + SS_{residual}. \quad (3)$$

Thus, R^2 can be rewritten as shown in Equation (4):

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \quad (4)$$

The *R* in Equation (5) is referred to as multiple correlation (Hair et al., 2006):

$$R = \sqrt{R^2}. \quad (5)$$

R^2 is referred to as squared multiple correlation or the coefficient of determination.

For an IV x_i in Equation (1), $i \in [1 \ldots j]$, with x_i as a DV and the remainings as IVs, a linear regression for determining the coefficient of determination of x_i is given by Equation (6):

$$x_i = a_1x_1 + a_2x_2 + \cdots + a_{i-1}x_{i-1} + a_{i+1}x_{i+1} + \cdots + a_jx_j + c_i,\ a_0x_0 = 0. \quad (6)$$

The coefficient of determination R_i^2 is calculated with x_i as the DV, then *tolerance* T_i and VIF_i are derived as Equation (7) and Equation (8) correspondingly:

$$T_i = 1 - R_i^2. \quad (7)$$

$$VIF_i = \frac{1}{T_i} \quad (8)$$

(B) Using *VIF*≥10 to Determine Whether Multicollinearity Exists

Although *VIF* has a clear mathematical definition, there is no fixed standard for this index. Indeed, standards may vary substantially. Nevertheless, many journal papers, postgraduate theses, and doctoral dissertations use *VIF*≥10 as an indicator of multicollinearity among IVs. Wen (2013) observed that from 2007, most of the papers in NTU Management Review used *VIF* to discuss multicollinearity, with *VIF*=10 as the cutoff value. Similarly, on analyzing Journal of Educational Research and Development from 2006 to 2020, five types of discussions on multicollinearity are found as follows:

1. describing the dangers of multicollinearity but not presenting the means and standards of diagnosing multicollinearity (Chang, 2012; Lin & Chien, 2019);

2. adopting *VIF* to diagnose multicollinearity but not clearly explaining the standard that was used (Lin & Tsai, 2014; Wang & Hsiao, 2006; Yeh, 2020);

3. adopting the correlation coefficients among IVs to diagnose multicollinearity but using different standards, including r≥.8 (Wu, 2020), *r*≥.9, and *r*≥.95 (Jeng & Chen, 2007);

4. using *VIF*≥10 to determine whether multicollinearity exists among the IVs (Chang, 2017; Chao & Luh, 2019; Lee & Yu, 2007);

5. adopting other methods such as removing variables or Partial Least Squares Structural Equation Modeling (PLS-SEM) to diagnose and inhibit multicollinearity but not specifying the judgment criteria (Chen, Li, & Tung, 2019; Wang, 2011).

International studies also frequently use the standard of *VIF*≥10 as a rule of thumb in multicollinearity detection (Cohen et al., 2003). Hair et al. (2006) confirmed that most studies use *VIF*≥10 to determine whether multicollinearity exists among IVs, despite the fact that this threshold still allows for high multicollinearity. For example, when multiple correlation *R* equals .9 (meaning that R^2 equals .81), the *tolerance* is .19 and the corresponding *VIF* is already 5.3 (Hair et al., 2006). In other words, if it only takes a *tolerance* of less than .19 and *VIF* greater than 5.3, then the correlation among the IVs will be greater than .9, which high level of correlation among IVs indicates multicollinearity in a linear regression. Consequently, Hair et al. pointed out that multicollinearity may even exist among IVs when *VIF* ranges from 3 to 5. However, many studies continue to use *VIF*=10 as a threshold and cite that this threshold is suggested by Hair et al. These studies might misinterpret what Hair et al.'s meaning. This is an issue worth exploring further.

(C) Adopting Different *VIF* Values for Multicollinearity Diagnostics

Although *VIF*≥10 is a common rule of thumb, a simple example can be used to demonstrate that this criterion is not reliable. In Table 1, *Y* is the DV and X_1 are X_2 are the IVs. The correlation coefficient of X_1 and X_2 is r=.904.

Table 1

Example of multicollinearity between IVs for *VIF*<10

Y	X_1	X_2
25	128	126
30	132	129
45	145	135
50	148	150
35	140	135
40	142	140
31	138	131
27	135	127
21	122	121
38	142	138

Regression of the individual IVs with regard to DV *Y* gives Y=1.076X_1-113.415 and Y=1.034X_2-103.515. The standardized regression coefficients corresponding to the two regression equations are Z_Y = .944Z_{X1} and Z_Y=.937Z_{X2}, respectively, indicating that the each of the IVs has significant predictive power with regard to DV *Y*. However, Table 2 represents the statistical attributes of the regression of the two IVs X_1 and X_2 with regard to dependent variable *Y*. Table 2 shows that neither of the two IVs has statistical significance. Clearly, collinearity exists between X_1 and X_2, causing them to drag each

other down and reducing the significance of both. It's worth noting that the *VIF* of X_1 and X_2 at this point is 5.459, which is far smaller than the commonly-applied cutoff value *VIF*=10. This example represents the first three of the four dangers discussed by Lewis-Beck (1980), and empirically corroborates Hair et al.'s suggestion.

Table 2

Linear regression analysis of example

Variable	Estimate	*SE*	95% CI		*p*	*VIF*
			LL	*UL*		
Intercept	-107.612	6.046	-121.907	-93.316	.000	
X_1	.073	.130	-.234	.379	.593	8.788
X_2	.984	.120	.701	1.267	.000	8.788

Some readers may argue that the above example is too simplistic, and especially the sample size is too small. To settle the argument, a simulation program listed in Appendix 1 initially provides as many as 5,000 normally distributed samples (in Line 7, 17, and 18) for each variable. The sample size of simulation can be altered by assigning different value to variable N in Line 7. Additionally, readers may set different value to variable corr in order to generate desire paired correlation coefficients of two IVs for a numerical simulation. From commented Line 10 to commented Line 16, the simulation program prepares statements for generating variables with three more different distributions such as binary, Poisson, and Gamma. Using the proposed program in Appendix 1 with different sample size, paired correlation coefficients, and distributions of IVs, the more general results are still closed to the outcomes of simplified example in Table 2 and therefore the discussion about the simplified example is valid.

The discussion for the simplified model in Table 1 shows that using *VIF*≥10 to determine whether multicollinearity exists among IVs may lead to misjudgment. Some researchers believe that multicollinearity may become a problem when *VIF* equals 4 or 5 (Pan & Jackson, 2008; Rogerson, 2001) or is between 3 and 5 (Hair et al., 2006). As seen in Equation (7) and Equation (8), the *tolerance* $T_i = 1 - R_i^2$ is the proportion of the variance in the IV x_i that is unexplained by the other IVs. Then $VIF_i = \frac{1}{T_i}$ can be explained as the magnification factor of total variance to the unexplained variance in Equation (6), and therefore $\sqrt{VIF}$ can be interpreted as the inflation times of standard error in a linear regression analysis. Because $\sqrt{VIF}$ is easier to interpret, it has been suggested that $\sqrt{VIF}$ can be used to observe the multicollinearity among IVs. For example, suppose that *VIF*=4 and $\sqrt{VIF}$=2, in this situation, the standard error of linear regression analysis is twice as high as when *VIF*=1. For the above reason, Miles & Shevlin (2001) suggest *VIF*≥4 as a criterion for determining whether multicollinearity exists.

Paired correlation coefficient *r* is often used as a tool for multicollinearity

diagnostics. Most studies adopt $r \geq .8$ as the cutoff (Vatcheva et al., 2016). However, these studies tend to use R^2 as the basis for judgment (Lewis-Beck, 1980). Based on the suggestions made by Lewis-Beck (1980) and Vatcheva et al. (2016), it can be inferred that if multiple correlation $R=.8$ is used as the cutoff of multicollinearity detection, then $R^2=.64$ and $T=.36$, which means that the cutoff value of *VIF* is 1/.36=2.778. As a result, some researchers believe that multicollinearity may become an issue if a *VIF* value is greater than 2 (Jeng, 2021; Sellin, 1990, as cited in Darmawan & Keeves, 2006).

(D) Influence of Number of IVs on VIF

Vatcheva et al. (2016) employed two and three IVs to discuss the relationship between multicollinearity and changes in paired correlation. The results of their experiments reveal that $VIF<5$ does not mean no multicollinearity exists among IVs. They then suggested that *VIF* judgment should be even more cautious when there are more IVs. They nonetheless did not present the relationship between the number of IVs and the *VIF*. However, theoretically, for a linear regression analysis equation with an infinite number of IVs, the R^2 of the equation will ultimately equal 1 (Berry, 1993). From this, it can be inferred that a greater number of IVs in a linear regression analysis equation makes R^2 grow faster and therefore multiple correlation R is more significant. The ascending of R^2 will have a knock-on effect on making *tolerance* smaller, and ultimately inflate the *VIF*. Therefore, the number of IVs should also be a crucial variable in determining the cutoff value of the *VIF*.

B. Research Motives and Objectives

The *VIF* criterion used to determine the multicollinearity among IVs varies from researcher to researcher. Although $VIF \geq 10$ is the most commonly used, it is not strict enough. Furthermore, although *VIF* is thought to be associated with the number of IVs, a relationship between the two has yet to be proposed. Thus, the objectives of this study are to determine the relationships among the number of IVs, correlations, and *VIF*, then to propose suitable values for *VIF* cutoffs.

II. Research Methods

In order to explore changes in the *VIF* resulting from different numbers of IVs and varying degrees of correlation among them, numerical simulations are introduced in this section.

A. Steps in the Simulation Process

The process of the numerical simulations is as follows:

Step 1. Import relevant libraries and initialize variable settings.

Step 2. Based on the paired correlations among the IVs, initially generate a normally-distributed dataset with 1,000 data for each variable. Note that the standard deviation of each variable is 1 and the mean of each variable equals 0.

Step 3. Calculate the *VIF* values of all of the IVs in Step 2. Due to the fact that the paired correlations r among the IVs are identical, the R^2 values of the IVs will all be the same, as will the *VIF* values. Repeat Step 2 and Step 3 until the set maximum correlation coefficient has been reached.

Step 4. Sort the correlation coefficients, numbers of IVs, and the corresponding *VIF* values and draw a graph.

B. Development Platform

This program was developed using the R platform Ver. 4.1.0. Its final version is shown in Appendix 2.

C. Explanation of the Simulation Program

Lines 1 through 6 in Appendix 2 import the libraries needed to develop the program. Table 3 presents the purposes and functions of the imported libraries. All libraries used are cited and referenced so that other researchers can install the libraries in order to run the program shown in Appendix 2. Lines 7 through 14 initialize the variables needed for the program, and Table 4 explains the meanings of the variables to help other researchers explore the results by changing the default values of variables.

Table 3

Purposes of libraries

Library name	Explanation of purpose
faux	Calls the rnorm_multi function and generates a normally-distributed dataset based on parameters such as mean, standard deviation, and correlation coefficients (DeBruine, 2021).
DAAG	Calls the vif function to calculate the *VIF* of each IV and forms a vector (Maindonald & Braun, 2020).
tibble	Calls the add_column function and adds a new column to the data frame (Müller & Wickham, 2021).
reshape2	Calls the melt function and converts the data frame into a dataset that can be drawn into a graph by ggplot (Wickham, 2007).
ggplot2	Calls the ggplot function and draws a statistical graph (Wickham, 2016).
GGally	An expanded library of ggplot2 (Schloerke et al., 2021).

Table 4

Names and purposes of variables

Variable name	Explanation of purpose
minNumIVs	Minimum number of IVs (there must be at least two IVs to check multicollinearity, so this variable was set at 2)
maxNumIVs	Maximum number of IVs
minCorr	Minimum value of correlation coefficients
maxCorr	Maximum value of correlation coefficients
corrStep	Step length from minCorr to maxCorr
N	Quantity of data to be generated for each IV
vecCorrelations	Vector from minCorr to maxCorr with corrStep as step length
rowCount	Number of rows needed to convert VIFVector into a data matrix

In Line 12, the default sample size 1,000 is assigned to variable N which is used to generate IVs in Line 18. The corr variable in the for loop beginning in Line 15 represents the correlation coefficients in the vector named vecCorrelations. In the loop, rnorm_multi is used to generate numVars groups of normally-distributed data with paired correlation as corr, mean as 0, and standard deviation as 1. When numVars = 6 and corr = .9, for instance, the datasets in Figure 1 are generated. As can be seen in Figure 1, the coefficients of the correlations between the variables equal .9, and the data within each variable present normal distributions. With Figure 1 as an example, the generated variables are sequentially renamed V1, V2, V3, …, V6, among which V1 is regarded as the DV and the remaining five variables, V2 to V6, are regarded as the IVs for *VIF* estimation. In

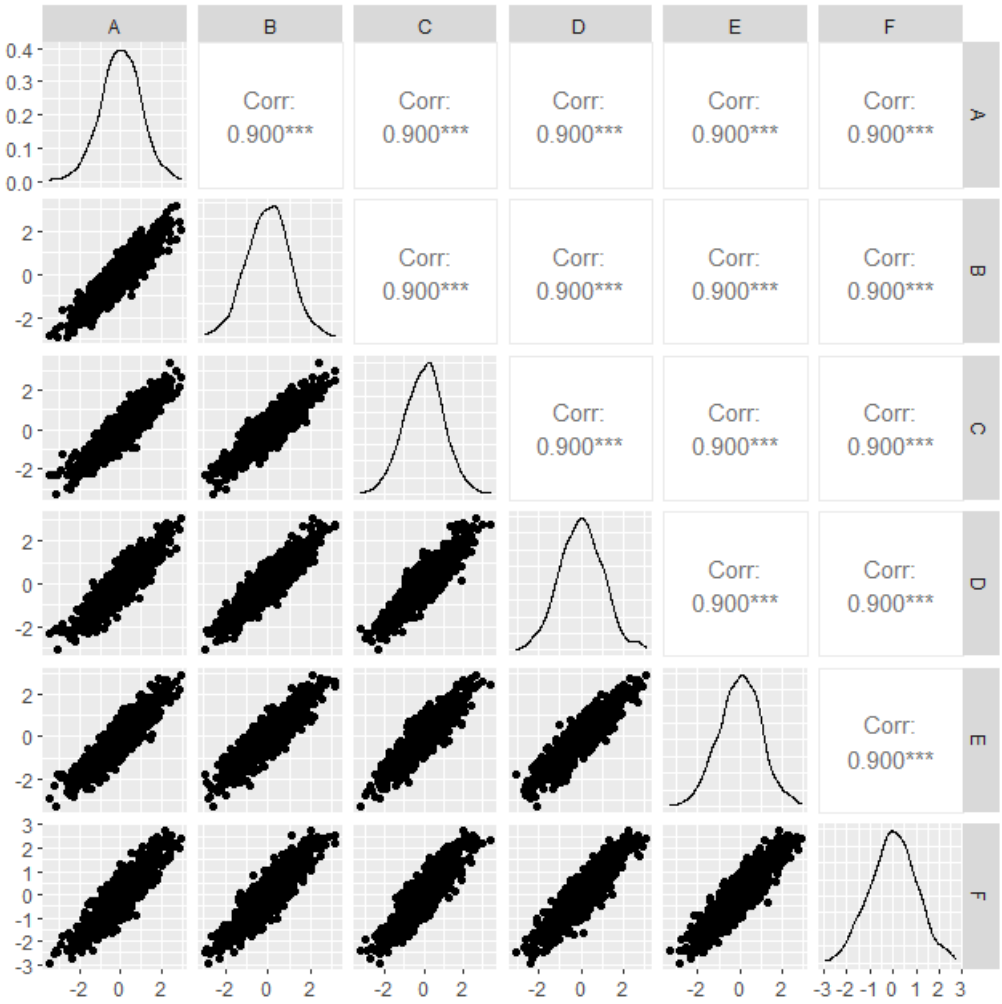
Line 22, a linear regression analysis model M is established to facilitate calculation of the *VIF* of model M in Line 23. Lines 24 through 28 connect the *VIF* values estimated from different correlation coefficients corr and different numbers of variables into a vector.

The purpose of Line 32 outside of the loop is to re-organize the aforementioned vector into the data matrix matrixVIF with rowCount rows and maxNumIVs-1 columns. Lines 33 to 37 convert the data matrix matrixVIF into the data frame format of R and name it VIFdata. Line 38 saves the data frame VIFdata as a csv file and names the csv file based on the maximum number of IVs and the correlation coefficient range of the data frame.

Lines 39 and 40 convert data frame VIFdata into a format compatible with ggplot function for graph drawing. Finally, Line 41 plots the correlation coefficients of different IVs with the number of IVs as the x axis and the *VIF* value as the y axis.

Figure 1

Example of normally distributed IV dataset with 6 IVs and $r = .90$.



III. Results and Discussion

This section first displays the *VIF* curves resulting from different number of IVs and correlation coefficients, and explains the characteristics of these *VIF* curves. Next, the relationships between the *VIF* curve graphs presented in this study and the *VIF* cutoff values suggested by previous research are discussed. Finally, at the end of this section, a novel table lookup method to determine the *VIF* cutoff values is proposed.

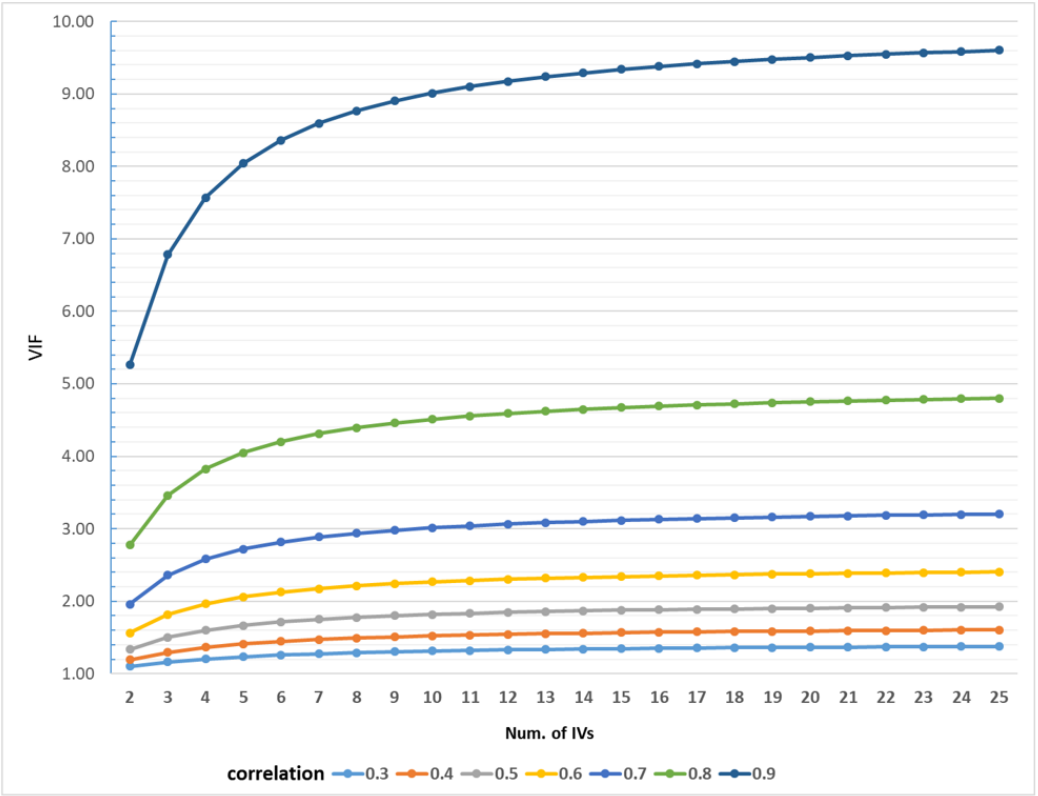
A. *VIF* Curves Corresponding to Different Numbers of IVs and Correlation Coefficients

Let the R program in Table 3 set maxNumIVs as 25, minCorr as .3, maxCorr as .9, and corrStep as .1. After the program with the above settings was executed, the curves were plotted as shown in Figure 2.

Observations of Figure 2 show the following: (1) The exponential rise to limits in *VIF* values when the correlation coefficients of the variables ranged from .3 to .9. (2) The intercept spacing between the curves of different correlation coefficients increases exponentially. For instance, the intercept spacing between the curves of correlation coefficients .9 and .8 is several times that between the curves of correlation coefficients .8 and .7. (3) *VIF* values increase with the number of independent IVs; however, the curves show that the *VIF* values of individual curves will reach a certain limit. For instance, the *VIF* limit of the curve of correlation coefficient .8 is around 5. (4) When the correlation coefficient of two variables has reached .9, the *VIF* is only slightly higher than 5.2. (5) Even with a correlation coefficient of .9 for 25 IVs, the *VIF* is still less than 10. (6) The *VIF* limits of individual curves with a correlation coefficient less than .9 among the IVs are all far less than 10.

Figure 2

VIF curves simulated with 1,000 samples for 2 to 25 IVs and *r* ranging from .3 to .9.

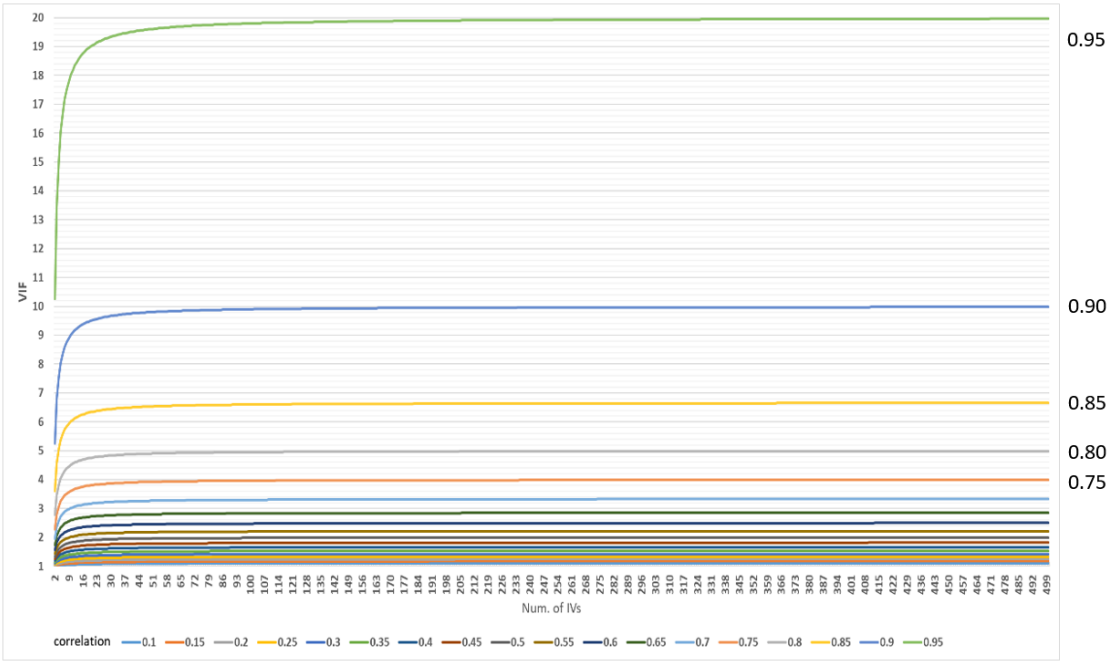


Next, setting maxNumIVs as 500, minCorr as .1, maxCorr as .95, and corrStep as .05, the program was executed to observe the curve trends with more IVs and expanded correlation coefficients from .1 to .95.

The trends of the curves in Figure 3 and the relationships among them are identical to those in Figure 2. However, it is worth noting that when the correlation coefficient is equal to .95 (top curve in Figure 3) and there are only 2 IVs, the *VIF* reaches 10.26, and when the number of IVs increases to 500, the *VIF* limit of the top curve in Fig 3 is close to 20.

Figure 3

VIF curves simulated with 1,000 samples for 2 to 500 IVs and *r* ranging from .1 to .95



B. Discussion

The results in Figs. 2 and 3 show that the changes in the *VIF* are associated with the correlation coefficients of the IVs and the number of IVs in regression models. This finding is similar to that of Vatcheva et al. (2016). However, this study further finds that the *VIF* does not increase infinitely with the number of IVs in regression models. When the coefficient of correlation among the IVs is a certain value, an increasing number of IVs will cause the *VIF* to converge at a certain limit.

It can also be noted that when the paired correlation of the IVs is .5 and the number of IVs increases, the *VIF* limit will be around 2. Furthermore, r=.5 is the median of [0,1], so if the correlation coefficient r of the IVs is less than .5, a low degree of correlation exists. When $r \leq .5$ and the number of IVs increases, the *VIF* will certainly be less than 2. It is worth noting that as shown in Figure 2, with a moderate degree of correlation (r=.6 or .7) and only two IVs, the *VIF* is still less than 2. Only when there is a sufficient number of IVs does the *VIF* exceed 2. Thus, if researchers believe that low degrees of correlation (r<.5) are important to their regression models, using $VIF \leq 2$ to detect multicollinearity might be useful. The discussion above corresponds to the suggestion made by some researchers to be wary of multicollinearity when *VIF* is around 2 (Jeng, 2021; Sellin, 1990).

A number of researchers have also suggested setting the cutoff value of *VIF* at 3, 4, or 5, with multicollinearity possible when *VIF* is between 3 and 5 (Hair et al, 2006; Miles & Shevlin, 2001; Pan & Jackson, 2008; Rogerson, 2001). The results of this study revealed that with an adequate number of IVs and correlation coefficients between .65

and .80, the value of the *VIF* should also be somewhere between 3 and 5. If there are relatively few IVs, such as two, and the correlation coefficients are between .8 and .9, then the *VIF* values will also fall between 2.78 and 5.26. When correlation coefficients are between .8 and .9 and so a high degree of correlation already exists among the IVs, the *VIF* values will still be around 3 to 5, which is remarkably lower than the cutoff value of *VIF*=10 suggested in most studies.

With regard to *VIF*≥10, observations of Figure 2 and 3 show that when the correlation coefficients of the IVs are .9, the *VIF* values are still less than 10 even when there are many IVs. From the two figures, it can be inferred that *VIF*=10 is the limit when *r*=.9. In other words, if *VIF*=10 serves as the cutoff value for multicollinearity, it will overlook the hazard of high correlation (*r*=.9) among the IVs. Further observation of Figure 3 revealed that *VIF*=10.26, which is only slightly greater than 10, when *r*=.95 and there are only two IVs. Researchers who are in the habit of rounding their numbers will find no multicollinearity even when a high degree of correlation exists among IVs (*r*=.95).

The discussion in this section shows that *VIF* values increase with the correlation coefficients of the IVs; they also increase with the number of IVs but tend to converge at a limit value. Two factors should therefore be taken into consideration when *VIF* is used to detect multicollinearity in a linear regression analysis: (1) the minimum degree of correlation among the IVs desired by the researchers based on the characteristics of their studies, and (2) the number of IVs in the linear regression analysis.

C. Method to Determine *VIF* Cutoffs

As changes in the *VIF* are associated with the paired correlations among IVs and the number of IVs, these two factors should both be taken into consideration when multicollinearity diagnostics are conducted during a linear regression analysis. Assigning 30 to maxNumIVs, .4 to minCorr, .99 to maxCorr, and .01 to corrStep, the program in Appendix 2 produced Table 6 and saved Table 6 as a csv file. In Table 6, the rows indicate the correlation coefficients (*r*) and the columns indicate the number of IVs (*k*). Two examples are given below to demonstrate how this table is used:

Example 1: Suppose a linear regression analysis in a study contains five IVs (*k*=5). The researchers believe the best degree of correlation among the IVs to be around .5 (*r*=.5). Looking up the row *r*=.5 and the column *k*=5 in Table 6 thus gives *VIF*=1.67 as the cutoff value in their study to determine whether the IVs in the linear regression model have issue of multicollinearity.

If the variables in the regression model are not paired correlated, the researchers may adopt a stepwise method to diagnose multicollinearity. In this case, a full model is first examined with the 5 IVs and then the IV which *VIF* is the highest, not significant, and is equal or greater than the initial cutoff value 1.67 is eliminated from the full model. Since there are 4 IVs left in the regression model, by looking up the row *r*=.5 and the column

k=4 in Table 6 the researchers obtain *VIF*=1.60 as the new cutoff value for the next round of stepwise regression. Repeating the process and deleting one IV at a time, in the end the final model retains the most uncorrelated IVs.

Example 2: Suppose a regression model has 7 IVs. A statistical analysis reveals that the *VIF* values of some of the IVs are greater than 3. Table 6 shows that when *k*=7 and *VIF* values are in interval between 3.0 and 4.0, the correlation coefficients of the IVs (*r*) may range from .72 to .78, which indicates moderately high correlation. Because high correlation coefficients between variables usually cause multicollinearity, in this case, the researchers should suspect that multicollinearity exists among the IVs.

IV. Conclusions, Limitations of the Study and Directions for Future Research

In this section, brief conclusions are drawn from the multicollinearity simulations, then the limitations of the study are summarized. Finally, the directions for future research are suggested.

A. Conclusions

Based on different considerations, researchers in the past have adopted varying *VIF* cutoff values, the most common of which is *VIF*=10. However, the numerical simulations proposed in this study reveal that *VIF*=10 is not strict enough as a cutoff value. It is also suggested that the number of IVs and the degree of correlation among the IVs must also be taken into account when determining a suitable *VIF* value. Accordingly, *VIF* cutoff values should be determined individually based on the number of IVs and the degree of correlation among the IVs. For a smaller correlation coefficient (*r*) or number of IVs (*k*), the *VIF* cutoff value should be more conservative. As a result, the table of *VIF* cutoff values proposed in this study is an improvement tool that does help researchers to look up the most appropriate cutoff values for multicollinearity diagnostics. Some researchers may consider using *tolerance* instead of *VIF*. Since *tolerance* is a reciprocal of *VIF*, researchers still can look up the Table 6 to obtain a *VIF* and then get a *tolerance* cutoff value by calculating the multiplicative inverse of the *VIF*. However, the direction of explanation for a *tolerance* cutoff should be reversed to its reciprocal value *VIF*.

It is also found that with a fixed correlation coefficient (*r*) and an increasing number of IVs (*k*), the *VIF* converges. However, the convergence is practically provided, but not mathematically proven. To put it simply, the *VIF* cutoff can be defined a function of correlation coefficient (*r*) and the number of IVs (*k*). This assumption will require further formal study.

B. Limitations of this Study and Suggestions for Future Research

There are several shortcomings of this study. First, the data are generated by simulation programs not from real life. Second, the *VIF* computations simplistically constraint on paired correlated IVs that does not completely correspond to the definition of multicollinearity. Third, the simulations only act on continuous data. Nevertheless, a researcher may have a linear regression model with different type of IVs: continuous, categorical, and ordinal predictors. Future research could examine the proposed method with large real-life sample and different data types of IVs. Because practical difficulties of simulating IVs with multiple correlations were encountered in this study, it is suggested that future researchers could develop relatively complete programs to examine multicollinearity of linear regression models with mixed-type variables.

According to Example 1 in section Method to Determine *VIF* Cutoffs, if the variables are not paired correlated, it is recommended to use the stepwise regression method for analysis. However, this method has its limitations. Stepwise regression is a data-driven approach in regression analysis that may result in retaining high-cost variables or deleting theoretically important variables when dealing with multicollinear variables. Therefore, it is recommended for researchers to consider using theory-driven hierarchical regression methods for addressing this issue. However, this study did not discuss or experiment with hierarchical regression methods for handling multicollinear variables. It is suggested that readers conduct further research in the future.

In the end, simulating the changes in *VIF* based on the paired correlation of IVs is simple and has been used in past studies to detect multicollinearity among IVs. However, multicollinearity means that an IV may have a multiple correlation rather than a paired correlation with other IVs. Despite this drawback, using paired correlation to estimate the changes in *VIF* produces the lower bound of *VIF* values and therefore still has applicable value.

Table 6

VIF cutoff values

k / *r*	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0.40	1.19	1.30	1.36	1.41	1.44	1.47	1.49	1.51	1.52	1.53	1.54	1.55	1.56	1.57	1.57	1.58	1.58	1.59	1.59	1.59	1.60	1.60	1.60	1.60	1.61	1.61	1.61	1.61	1.61
0.41	1.20	1.31	1.38	1.43	1.47	1.49	1.52	1.53	1.55	1.56	1.57	1.58	1.59	1.59	1.60	1.60	1.61	1.61	1.62	1.62	1.62	1.63	1.63	1.63	1.63	1.64	1.64	1.64	1.64
0.42	1.21	1.33	1.40	1.45	1.49	1.52	1.54	1.56	1.57	1.58	1.60	1.60	1.61	1.62	1.62	1.63	1.64	1.64	1.64	1.65	1.65	1.65	1.66	1.66	1.66	1.66	1.67	1.67	1.67
0.43	1.23	1.35	1.43	1.48	1.51	1.54	1.57	1.58	1.60	1.61	1.62	1.63	1.64	1.65	1.65	1.66	1.66	1.67	1.67	1.68	1.68	1.68	1.69	1.69	1.69	1.69	1.69	1.70	1.70
0.44	1.24	1.37	1.45	1.50	1.54	1.57	1.59	1.61	1.63	1.64	1.65	1.66	1.67	1.68	1.68	1.69	1.69	1.70	1.70	1.71	1.71	1.71	1.72	1.72	1.72	1.72	1.72	1.73	1.73
0.45	1.25	1.39	1.47	1.53	1.57	1.60	1.62	1.64	1.66	1.67	1.68	1.69	1.70	1.71	1.71	1.72	1.72	1.73	1.73	1.74	1.74	1.74	1.75	1.75	1.75	1.75	1.76	1.76	1.76
0.46	1.27	1.41	1.49	1.55	1.59	1.63	1.65	1.67	1.69	1.70	1.71	1.72	1.73	1.74	1.74	1.75	1.76	1.76	1.76	1.77	1.77	1.78	1.78	1.78	1.78	1.79	1.79	1.79	1.79
0.47	1.28	1.43	1.52	1.58	1.62	1.65	1.68	1.70	1.72	1.73	1.74	1.75	1.76	1.77	1.78	1.78	1.79	1.79	1.80	1.80	1.81	1.81	1.81	1.81	1.82	1.82	1.82	1.82	1.83
0.48	1.30	1.45	1.54	1.61	1.65	1.69	1.71	1.73	1.75	1.76	1.78	1.79	1.80	1.80	1.81	1.82	1.82	1.83	1.83	1.84	1.84	1.84	1.85	1.85	1.85	1.85	1.86	1.86	1.86
0.49	1.32	1.48	1.57	1.64	1.68	1.72	1.74	1.77	1.78	1.80	1.81	1.82	1.83	1.84	1.85	1.85	1.86	1.86	1.87	1.87	1.88	1.88	1.88	1.89	1.89	1.89	1.89	1.90	1.90
0.50	1.33	1.50	1.60	1.67	1.71	1.75	1.78	1.80	1.82	1.83	1.85	1.86	1.87	1.88	1.88	1.89	1.89	1.90	1.90	1.91	1.91	1.92	1.92	1.92	1.93	1.93	1.93	1.93	1.94
0.51	1.35	1.53	1.63	1.70	1.75	1.78	1.81	1.84	1.85	1.87	1.88	1.89	1.90	1.91	1.92	1.93	1.93	1.94	1.94	1.95	1.95	1.96	1.96	1.96	1.97	1.97	1.97	1.97	1.97
0.52	1.37	1.55	1.66	1.73	1.78	1.82	1.85	1.87	1.89	1.91	1.92	1.93	1.94	1.95	1.96	1.97	1.97	1.98	1.98	1.99	1.99	2.00	2.00	2.00	2.01	2.01	2.01	2.01	2.02
0.53	1.39	1.58	1.69	1.77	1.82	1.86	1.89	1.91	1.93	1.95	1.96	1.97	1.98	1.99	2.00	2.01	2.02	2.02	2.03	2.03	2.03	2.04	2.04	2.05	2.05	2.05	2.05	2.06	2.06
0.54	1.41	1.61	1.73	1.80	1.86	1.90	1.93	1.95	1.97	1.99	2.00	2.02	2.03	2.04	2.04	2.05	2.06	2.06	2.07	2.07	2.08	2.08	2.09	2.09	2.09	2.10	2.10	2.10	2.10
0.55	1.43	1.64	1.76	1.84	1.90	1.94	1.97	2.00	2.02	2.03	2.05	2.06	2.07	2.08	2.09	2.10	2.10	2.11	2.12	2.12	2.12	2.13	2.13	2.14	2.14	2.14	2.15	2.15	2.15
0.56	1.46	1.67	1.80	1.88	1.94	1.98	2.01	2.04	2.06	2.08	2.10	2.11	2.12	2.13	2.14	2.14	2.15	2.16	2.16	2.17	2.17	2.18	2.18	2.18	2.19	2.19	2.19	2.20	2.20
0.57	1.48	1.71	1.84	1.92	1.98	2.03	2.06	2.09	2.11	2.13	2.14	2.16	2.17	2.18	2.19	2.19	2.20	2.21	2.21	2.22	2.22	2.23	2.23	2.24	2.24	2.24	2.24	2.25	2.25

Table 6 (continued)

VIF cutoff values

r \ *k*	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0.58	1.51	1.74	1.88	1.97	2.03	2.07	2.11	2.14	2.16	2.18	2.19	2.21	2.22	2.23	2.24	2.25	2.25	2.26	2.27	2.27	2.28	2.28	2.28	2.29	2.29	2.30	2.30	2.30	2.30
0.59	1.53	1.78	1.92	2.01	2.07	2.12	2.16	2.19	2.21	2.23	2.25	2.26	2.27	2.28	2.29	2.30	2.31	2.32	2.32	2.33	2.33	2.34	2.34	2.34	2.35	2.35	2.35	2.36	2.36
0.60	1.56	1.82	1.96	2.06	2.13	2.17	2.21	2.24	2.27	2.29	2.30	2.32	2.33	2.34	2.35	2.36	2.37	2.37	2.38	2.38	2.39	2.39	2.40	2.40	2.41	2.41	2.41	2.42	2.42
0.61	1.59	1.86	2.01	2.11	2.18	2.23	2.27	2.30	2.32	2.34	2.36	2.38	2.39	2.40	2.41	2.42	2.43	2.43	2.44	2.45	2.45	2.46	2.46	2.46	2.47	2.47	2.47	2.48	2.48
0.62	1.62	1.90	2.06	2.16	2.23	2.29	2.33	2.36	2.38	2.41	2.42	2.44	2.45	2.46	2.47	2.48	2.49	2.50	2.50	2.51	2.52	2.52	2.52	2.53	2.53	2.54	2.54	2.54	2.55
0.63	1.66	1.95	2.11	2.22	2.29	2.35	2.39	2.42	2.45	2.47	2.49	2.50	2.52	2.53	2.54	2.55	2.56	2.56	2.57	2.58	2.58	2.59	2.59	2.60	2.60	2.60	2.61	2.61	2.61
0.64	1.69	2.00	2.17	2.28	2.35	2.41	2.45	2.49	2.51	2.54	2.56	2.57	2.59	2.60	2.61	2.62	2.63	2.64	2.64	2.65	2.65	2.66	2.66	2.67	2.67	2.68	2.68	2.68	2.69
0.65	1.73	2.05	2.23	2.34	2.42	2.48	2.52	2.56	2.59	2.61	2.63	2.65	2.66	2.67	2.68	2.69	2.70	2.71	2.72	2.72	2.73	2.74	2.74	2.75	2.75	2.75	2.76	2.76	2.76
0.66	1.77	2.10	2.29	2.41	2.49	2.55	2.60	2.63	2.66	2.69	2.71	2.72	2.74	2.75	2.76	2.77	2.78	2.79	2.80	2.80	2.81	2.82	2.82	2.83	2.83	2.83	2.84	2.84	2.84
0.67	1.81	2.16	2.36	2.48	2.56	2.63	2.67	2.71	2.74	2.77	2.79	2.81	2.82	2.83	2.85	2.86	2.87	2.87	2.88	2.89	2.90	2.90	2.91	2.91	2.92	2.92	2.92	2.93	2.93
0.68	1.86	2.22	2.43	2.55	2.64	2.71	2.76	2.80	2.83	2.85	2.87	2.89	2.91	2.92	2.94	2.95	2.96	2.96	2.97	2.98	2.99	2.99	3.00	3.00	3.01	3.01	3.02	3.02	3.02
0.69	1.91	2.29	2.50	2.63	2.73	2.79	2.84	2.88	2.92	2.94	2.97	2.99	3.00	3.02	3.03	3.04	3.05	3.06	3.07	3.08	3.08	3.09	3.09	3.10	3.10	3.11	3.11	3.12	3.12
0.70	1.96	2.36	2.58	2.72	2.81	2.88	2.94	2.98	3.01	3.04	3.07	3.09	3.10	3.12	3.13	3.14	3.15	3.16	3.17	3.18	3.18	3.19	3.20	3.20	3.21	3.21	3.22	3.22	3.22
0.71	2.02	2.44	2.67	2.81	2.91	2.98	3.04	3.08	3.12	3.15	3.17	3.19	3.21	3.22	3.24	3.25	3.26	3.27	3.28	3.29	3.29	3.30	3.31	3.31	3.32	3.32	3.33	3.33	3.33
0.72	2.08	2.52	2.76	2.91	3.01	3.09	3.15	3.19	3.23	3.26	3.28	3.30	3.32	3.34	3.35	3.37	3.38	3.39	3.40	3.40	3.41	3.42	3.43	3.43	3.44	3.44	3.45	3.45	3.45
0.73	2.14	2.60	2.86	3.01	3.12	3.20	3.26	3.31	3.35	3.38	3.40	3.43	3.45	3.46	3.48	3.49	3.50	3.51	3.52	3.53	3.54	3.55	3.55	3.56	3.56	3.57	3.57	3.58	3.58
0.74	2.21	2.70	2.96	3.13	3.24	3.32	3.39	3.43	3.47	3.51	3.53	3.56	3.58	3.60	3.61	3.62	3.64	3.65	3.66	3.67	3.67	3.68	3.69	3.69	3.70	3.71	3.71	3.72	3.72
0.75	2.29	2.80	3.08	3.25	3.37	3.45	3.52	3.57	3.61	3.65	3.68	3.70	3.72	3.74	3.76	3.77	3.78	3.79	3.80	3.81	3.82	3.83	3.84	3.84	3.85	3.85	3.86	3.86	3.87

Table 6 (continued)

VIF cutoff values

k / *r*	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0.76	2.37	2.91	3.20	3.38	3.51	3.60	3.67	3.72	3.76	3.80	3.83	3.85	3.88	3.89	3.91	3.93	3.94	3.95	3.96	3.97	3.98	3.99	4.00	4.00	4.01	4.01	4.02	4.02	4.03
0.77	2.46	3.03	3.34	3.53	3.66	3.75	3.82	3.88	3.93	3.96	3.99	4.02	4.04	4.06	4.08	4.10	4.11	4.12	4.13	4.14	4.15	4.16	4.17	4.18	4.18	4.19	4.19	4.20	4.20
0.78	2.55	3.16	3.48	3.68	3.82	3.92	4.00	4.06	4.10	4.14	4.18	4.20	4.23	4.25	4.27	4.28	4.30	4.31	4.32	4.33	4.34	4.35	4.36	4.37	4.37	4.38	4.38	4.39	4.40
0.79	2.66	3.30	3.65	3.86	4.00	4.11	4.19	4.25	4.30	4.34	4.37	4.40	4.43	4.45	4.47	4.49	4.50	4.51	4.53	4.54	4.55	4.56	4.57	4.57	4.58	4.59	4.59	4.60	4.60
0.80	2.78	3.46	3.82	4.05	4.20	4.31	4.39	4.46	4.51	4.56	4.59	4.62	4.65	4.67	4.69	4.71	4.73	4.74	4.75	4.76	4.78	4.78	4.79	4.80	4.81	4.82	4.82	4.83	4.83
0.81	2.91	3.64	4.02	4.26	4.42	4.54	4.62	4.69	4.75	4.79	4.83	4.87	4.89	4.92	4.94	4.96	4.97	4.99	5.00	5.02	5.03	5.04	5.05	5.05	5.06	5.07	5.08	5.08	5.09
0.82	3.05	3.83	4.24	4.49	4.66	4.79	4.88	4.95	5.01	5.06	5.10	5.14	5.16	5.19	5.21	5.23	5.25	5.27	5.28	5.29	5.31	5.32	5.33	5.34	5.34	5.35	5.36	5.37	5.37
0.83	3.21	4.05	4.48	4.75	4.93	5.07	5.17	5.24	5.31	5.36	5.40	5.44	5.47	5.50	5.52	5.54	5.56	5.58	5.59	5.60	5.62	5.63	5.64	5.65	5.66	5.67	5.67	5.68	5.69
0.84	3.40	4.29	4.76	5.05	5.24	5.38	5.49	5.57	5.64	5.69	5.74	5.78	5.81	5.84	5.86	5.89	5.91	5.92	5.94	5.96	5.97	5.98	5.99	6.00	6.01	6.02	6.03	6.04	6.04
0.85	3.60	4.57	5.07	5.38	5.59	5.74	5.85	5.94	6.01	6.07	6.12	6.16	6.20	6.23	6.25	6.28	6.30	6.32	6.34	6.35	6.37	6.38	6.39	6.40	6.41	6.42	6.43	6.44	6.45
0.86	3.84	4.88	5.43	5.76	5.98	6.15	6.27	6.36	6.44	6.50	6.56	6.60	6.64	6.67	6.70	6.73	6.75	6.77	6.79	6.81	6.82	6.83	6.85	6.86	6.87	6.88	6.89	6.90	6.91
0.87	4.11	5.25	5.84	6.20	6.44	6.62	6.75	6.85	6.93	7.00	7.06	7.11	7.15	7.18	7.22	7.24	7.27	7.29	7.31	7.33	7.35	7.36	7.37	7.39	7.40	7.41	7.42	7.43	7.44
0.88	4.43	5.68	6.32	6.71	6.98	7.17	7.31	7.42	7.51	7.59	7.65	7.70	7.74	7.78	7.82	7.85	7.87	7.90	7.92	7.94	7.96	7.97	7.99	8.00	8.01	8.03	8.04	8.05	8.06
0.89	4.81	6.18	6.89	7.32	7.61	7.81	7.97	8.09	8.19	8.27	8.34	8.40	8.45	8.49	8.53	8.56	8.59	8.62	8.64	8.66	8.68	8.70	8.71	8.73	8.74	8.76	8.77	8.78	8.79
0.90	5.26	6.79	7.57	8.04	8.36	8.59	8.77	8.90	9.01	9.10	9.17	9.24	9.29	9.34	9.38	9.42	9.45	9.48	9.50	9.53	9.55	9.57	9.59	9.60	9.62	9.63	9.64	9.66	9.67
0.91	5.82	7.53	8.40	8.93	9.29	9.55	9.74	9.89	10.01	10.11	10.19	10.26	10.32	10.38	10.42	10.46	10.50	10.53	10.56	10.58	10.61	10.63	10.65	10.67	10.69	10.70	10.72	10.73	10.74
0.92	6.51	8.45	9.44	10.04	10.45	10.74	10.95	11.12	11.26	11.37	11.47	11.55	11.61	11.67	11.72	11.77	11.81	11.85	11.88	11.91	11.93	11.96	11.98	12.00	12.02	12.04	12.06	12.07	12.09
0.93	7.40	9.64	10.78	11.47	11.93	12.27	12.52	12.71	12.87	13.00	13.10	13.19	13.27	13.34	13.40	13.45	13.50	13.54	13.57	13.61	13.64	13.67	13.69	13.72	13.74	13.76	13.78	13.79	13.81

Table 6 (continued)

VIF cutoff values

r \ *k*	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0.94	8.59	11.23	12.57	13.38	13.92	14.31	14.60	14.83	15.01	15.16	15.29	15.39	15.48	15.56	15.63	15.69	15.74	15.79	15.84	15.88	15.91	15.94	15.97	16.00	16.03	16.05	16.07	16.09	16.11
0.95	10.26	13.45	15.07	16.04	16.70	17.16	17.52	17.79	18.01	18.19	18.34	18.47	18.58	18.67	18.75	18.83	18.89	18.95	19.00	19.05	19.09	19.13	19.17	19.20	19.23	19.26	19.29	19.31	19.34
0.96	12.76	16.78	18.81	20.04	20.86	21.45	21.89	22.24	22.51	22.74	22.92	23.08	23.22	23.34	23.44	23.53	23.61	23.69	23.75	23.81	23.87	23.92	23.96	24.00	24.04	24.08	24.11	24.14	24.17
0.97	16.92	22.34	25.06	26.71	27.81	28.59	29.18	29.64	30.01	30.31	30.56	30.78	30.96	31.12	31.25	31.38	31.49	31.58	31.67	31.75	31.82	31.89	31.95	32.00	32.05	32.10	32.14	32.19	32.22
0.98	25.25	33.45	37.56	40.04	41.70	42.88	43.77	44.46	45.01	45.46	45.84	46.16	46.43	46.67	46.88	47.06	47.23	47.37	47.50	47.62	47.73	47.83	47.92	48.00	48.08	48.15	48.22	48.28	48.33
0.99	50.25	66.78	75.06	80.04	83.36	85.74	87.52	88.90	90.01	90.92	91.67	92.31	92.86	93.34	93.75	94.12	94.45	94.74	95.00	95.24	95.46	95.65	95.84	96.00	96.16	96.30	96.43	96.55	96.67

Appendix 1

Program for Examining Multicollinearity of Two IVs with Paired Correlation

Line no.	Statement
1	require(DAAG) #for VIF function
2	require(ggplot2) #for ggplot
3	require(GGally) #Extension to ggplot2
4	require(faux) #for rnorm_multi
5	require(simstudy) #for genCorGen
6	require(lawstat) #for Levene test
7	N <- 5000 #sample size
8	corr <- .904 #paired correlation
9	numVars <- 3 #1 for dependent variable and 2 for independent variables
10	#l <- c(.91, .91, .91) # lambda for each new variable
11	#dat <- genCorGen(N, nvars = 3, params1 = l, dist = "binary", rho = .99,
	corstr = "cs", wide = TRUE) #Binary distribution
13	#l <- c(1, 1, 1) # lambda for each new variable
14	#dat <- genCorGen(N, nvars = 3, params1 = l, dist = "poisson", rho = .92,
	corstr = "cs", wide = TRUE) #Poisson distribution
15	#l <- c(1, 1, 1) # lambda for each new variable
16	#dat <- genCorGen(N, nvars = 3, params1 = l, params2 = c(1, 1, 1), dist =
	"gamma", rho = .92, corstr = "cs", wide = TRUE) #Gamma
	distribution
17	l <- c(1, 1, 1) # lambda for each new variable
18	dat <- genCorGen(N, nvars = 3, params1 = l, params2 = l, dist = "normal",
	rho = .95, corstr = "cs", wide = TRUE) #Normal distribution
19	cor(dat)
20	group <- rep(1:2, each=N)
21	values <- c(dat$V2, dat$V3)
22	groupeddat <- data.frame(group, values)
23	write.csv(groupeddat, "D:\\groupeddat.csv", row.names = TRUE)
24	#normality test
25	shapiro.test(dat$V2)
26	shapiro.test(dat$V3)
27	#linearity test
28	plot(dat$V2, dat$V3)
29	abline(lm(dat$V2 ~ dat$V3))
30	#variance homogeneity test
31	var.test(values ~ group, data = groupeddat)
32	levene.test(groupeddat$values, groupeddat$group)
33	#establish a linear regression model
34	print(ggpairs(dat))
35	M <- lm(V1~V2+V3, data=dat)
36	VIF <- vif(M)

Appendix 2

Program for analysis of VIF values corresponding to different numbers of IVs and paired correlations.

```
require(faux)              #for rnorm_multi
require(DAAG)              #for vif
require(tibble)            #for add_column
require(reshape2)          #for melt
require(ggplot2)           #for ggplot
require(GGally)            #extension to ggplot2
minNumIVs <- 2
maxNumIVs <- 30
minCorr <- 0.4
maxCorr <- 0.99
corrStep <- 0.01
N <- 1000
vecCorrelations <- seq(minCorr, maxCorr, by=corrStep)
rowCount <- 0
for (corr in vecCorrelations) {
    rowCount <- rowCount + 1
    for (numVars in (minNumIVs +1):(maxNumIVs +1)) {
        dat <- rnorm_multi(n = N,
                vars = numVars,
                mu = 0,
                sd = 1,
                r = corr,
                varnames = paste("V", seq(from=1, to=numVars), sep=""),
                empirical = TRUE)
        #cor(dat)
        #dev.new()
        #print(ggpairs(dat))
        M <- lm(V1~., data=dat)
        VIF <- mean(vif(M))
        if (corr == minCorr && numVars ==(minNumIVs +1)) {
            VIFvector <- VIF
        } else {
            VIFvector <- append(VIFvector, VIF)
        }
        #write.csv(dat, paste(numVars, "v_cor", corr, ".csv", sep=""))
    }
}
matrixVIF = matrix(VIFvector, nrow=rowCount, ncol=maxNumIVs -1,
            byrow=TRUE)
colnames <-    paste("", seq(from=minNumIVs , to=maxNumIVs ), sep="")
colnames <- c("corr", colnames)
VIFdata <- as.data.frame(matrixVIF)
VIFdata <- add_column(VIFdata, vecCorrelations, .before = 1)
colnames(VIFdata) <- colnames
write.csv(VIFdata, paste("2to",maxNumIVs , "v_cor", minCorr, "to",
```

```
38              maxCorr, "-VIF.csv", sep=""))
    melted = melt(VIFdata, id.vars="corr")
39  colnames(melted) <- c("vecCorrelations", "variables", "VIF")
40  ggplot(data=melted, aes(x=variables, y=VIF, group=vecCorrelations,
41          color=factor(vecCorrelations))) + scale_x_discrete(name="Num. of
            Variables", breaks=seq(0, maxNumIVs ,2))+
            scale_y_continuous(name="VIF", breaks=seq(0, 20, 0.5)) +
            geom_line()
```

References

Berry, W. D. (1993). *Understanding regression assumptions*. Sage Publications.

Chang, F. –C. (2012). Test of the relationships among the educational, economic development and political democracy: Time series exploration. *Journal of Educational Research and Development, 8*(3), 123-162. https://doi.org/10.6925/SCJ.201209_8(3).0005

[張芳全（2012）。教育、經濟發展與政治民主之關聯研究：跨時間數列的探索。**教育研究與發展期刊，8**（3），123-162。https://doi.org/10.6925/SCJ.201209_8(3).0005]

Chang, F. –C. (2017). A cross-nations study of the well-being and academic achievement. *Journal of Educational Research and Development, 13*(3), 31-66. https://doi.org/10.3966/181665042017091303002

[張芳全（2017）。幸福感與學習成就之跨國分析。**教育研究與發展期刊，13**（3），31-66。https://doi.org/10.3966/181665042017091303002]

Chao, T. –Y., & Luh, W. –M. (2019). The effects of over-education and mismatch on job satisfaction. *Journal of Educational Research and Development, 15*(3), 93-128. https://doi.org/10.3966/181665042019091503004

[趙子揚、陸偉明（2019）。過量教育、學用不一對於工作滿意度的影響。**教育研究與發展期刊，15**（3），93-128。https://doi.org/10.3966/181665042019091503004]

Chen, C. –Y., Li, C. –C., & Tung, Y. –Y. (2019). The effects of social strains on the illicit drugs prevention and the attitude toward illicit drugs use for high school students in Taiwan. *Journal of Educational Research and Development, 15*(4), 35-66. https://doi.org/10.3966/181665042019121504002

[陳芝吟、李承傑、董旭英（2019）。社會緊張對於高中職生毒品防治宣導與毒品使用態度之影響。**教育研究與發展期刊，15**（4），35-66。https://doi.org/10.3966/181665042019121504002]

Chiou, H. (2021). *Quantitative research methods II: Statistical principles and analytic techniques*. Yeh Yeh Book Gallery.

[邱皓政（2021）。**量化研究法(二)：統計原理與分析技術**。雙葉書廊。]

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Erlbaum.

Darmawan, I. G. H, & Keeves, J. P. (2006). Suppressor variables and multilevel mixture modelling. *International Education Journal, 7*(2), 160-173.

DeBruine, L. (2021). *faux: Simulation for factorial designs* (R package version 1.0.0) [Computer software]. The Comprehensive R Archive Network. https://debruine.github.io/faux/

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6 Ed.). Pearson Education.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer.

Jeng, C. –C. (2021). Hierarchical linear model approach to explore interaction effects of swimmers' characteristics and breathing patterns on swimming performance in butterfly stroke with self-developed inertial measurement unit. *Journal of Computers, 32*(3), 99-118. https://doi.org/10.3966/199115992021063203008

Jeng, H. –L., & Chen, S. –Y. (2007). Model exploration and validation of the standardized spatial ability test. *Journal of Educational Research and Development, 3*(4), 181-216.

[鄭海蓮、陳世玉（2007）。標準化空間能力測驗之建模與驗證。**教育研究與發展期刊，3**（4），181-216。]

Lee, D. R., & Yu, M. –N. (2007). The assessment of knowledge structures of learning performance in educational statistics. *Journal of Educational Research and Development, 3*(4), 113-148.

[李敦仁、余民寧（2007）。學習表現的知識結構評量研究：以「教育統計學」學科知識為例。**教育研究與發展期刊，3**（4），113-148。]

Lewis-Beck, M. S. (1980). *Applied regression: An introduction.* Sage Publications.

Lin, C. –S. (2014). *Psychological and educational statistics.* DongHua.

[林清山（2014）。**心理與教育統計學**。東華。]

Lin, H. –C., & Chien, W. –C. (2019). A study on learning cultural capital of disadvantaged students during the summer vacation in urban elementary schools in Taiwan. *Journal of Educational Research and Development, 15*(3), 23-58. https://doi.org/10.3966/181665042019091503002

[林信志、簡瑋成（2019）。臺灣都會地區國小弱勢學生暑期學習活動資本之研究。**教育研究與發展期刊，15**（3），23-58。https://doi.org/10.3966/181665042019091503002]

Lin, P. –Y., & Tsai, Y. –T. (2014). Kindergarten literacy environment and first graders' reading performance. *Journal of Educational Research and Development, 10*(2), 1-30. https://doi.org/10.3966/181665042014061002001

[林佩伃、蔡邑庭（2014）。小一學童的基礎閱讀表現與幼兒園閱讀素養環境的關聯。**教育研究與發展期刊，10**（2），1-30。https://doi.org/10.3966/181665042014061002001]

Maindonald, J. H., & Braun, W. J. (2020). *DAAG: Data analysis and graphics data and functions* (R package version 1.24) [Computer software]. The Comprehensive R Archive Network. https://CRAN.R-project.org/package=DAAG

Miles, J., & Shevlin, M. (2001). *Applying regression & correlation: A guide for students and researchers*. Sage Publications.

Müller, K., & Wickham, H. (2021). *tibble: Simple data frames* (R package version 3.1.3) [Computer software]. The Comprehensive R Archive. https://CRAN.R-project.org/package=tibble

Pan, Y., & Jackson, R. T. (2008). Ethnic difference in the relationship between acute inflammation and serum ferritin in US adult males. *Epidemiology and Infection, 136*(3), 421-431. https://doi.org/10.1017/S095026880700831X

Rogerson, P. A. (2001). *Statistical methods for geography*. Sage Publication.

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., & Crowley, J. (2021). *GGally: Extension to 'ggplot2'* (R package version 2.1.2) [Computer software]. The Comprehensive R Archive Network. https://CRAN.R-project.org/package=GGally

Sellin, N. (1990). *PLSPATH version 3.01: Program manual*. Universität Hamburg.

Vatcheva, K. P., Lee, M., McCormick, J. B., & Rahbar, M. H. (2016). Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology (Sunnyvale), 6*(2), 227. https://doi.org/10.4172/2161-1165.1000227

Wang, C. –H., & Hsiao, C. –T. (2006). An aptitude test to assess gifted and talented students. *Journal of Educational Research and Development, 2*(4), 143-162.

[王金香、蕭金土（2006）。**團體性向測驗在資賦優異學生甄試運用之研究。教育研究與發展期刊，2**（4），143-162。]

Wang, Y. –C. (2011). A six-step standardized model development strategy for hierarchical linear model. *Journal of Educational Research and Development, 7*(4), 25-55. https://doi.org/10.6925/SCJ.201112_7(4).0002

[王郁琮（2011）。**階層線性模式路徑圖與策略化模型建構機制。教育研究與發展期刊，7**（4），25-55。https://doi.org/10.6925/SCJ.201112_7(4).0002]

Wen, F. –H. (2013). Five important concepts of using regression analysis in social science studies. *Journal of Management, 30*(2), 169-190. https://doi.org/10.6504/JOM.2013.30.02.04

[溫福星（2013）。社會科學研究中使用迴歸分析的五個重要概念。**管理學報，30**（2），169-190。https://doi.org/10.6504/JOM.2013.30.02.04]

Wickham, H. (2007). *Reshaping data with the {reshape} package* [Computer software]. The Comprehensive R Archive Network. http://www.jstatsoft.org/v21/i12/

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* [Computer software]. The Comprehensive R Archive Network. https://ggplot2.tidyverse.org

Wu, C. –T. (2020). The impact of school knowledge governance on teachers' knowledge sharing and knowledge transfer: The mediating effect of social capital. *Journal of Educational Research and Development, 16*(3), 31-65. https://doi.org/10.6925/SCJ.202009_16(3).0002

[吳政達（2020）。學校知識治理對教師知識分享與知識轉移之影響：社會資本的中介效果分析。**教育研究與發展期刊，16**（3），31-65。https://doi.org/10.6925/SCJ.202009_16(3).0002]

Yeh, L. –C. (2020). Using IPMA-RG in educational research. *Journal of Educational Research and Development, 16*(4), 71-108. https://doi.org/10.6925/SCJ.202012_16(4).0003

[葉連祺（2020）。IPMA-RG 在教育研究之應用。**教育研究與發展期刊，16**（4），71-108。https://doi.org/10.6925/SCJ.202012_16(4).0003]

Yen, Y. –C. (1994). *Applied statistics*. Chao-Ming Chen.
[顏月珠（1994）。**實用統計方法(修訂本)：圖解與實例**。陳昭明。]

投稿日期：2022 年 08 月 25 日
修正日期：2023 年 07 月 01 日
接受日期：2023 年 09 月 22 日

為何多重共線性診斷不宜採用 VIF=10 為決斷值

鄭承昌*

中文摘要

進行迴歸分析時，常用變異數膨脹因子（Variance Inflation Factor, VIF）當作自變項是否具共線性的判別準則。過去研究者所建議的 *VIF* 決斷值，從 2 到 10 皆有，標準並不一致。本研究採用 R 模擬，針對不同數量的自變項（*n*）及自變項間的相關係數（*r*）計算 *VIF*，得以下結果：（1）*VIF* 和自變項的個數以及自變項間的相關係數有關；（2）當自變項間的相關係數為固定值，且自變項的個數趨近於無窮多時，*VIF* 會趨近於一特定的極限值；（3）共線性的判別不應該採行固定的 *VIF* 值為決斷值，應考量自變項個數及所能容忍的相關係數來判斷。基於結果，本研究提出 *VIF* 決斷值表，協助研究者能依據其迴歸分析中自變項個數及自變項相關程度來尋找 *VIF* 的決斷標準。最後，文末提出本研究的限制與建議。

關鍵詞：線性迴歸、多重共線性、變異數膨脹因子

*國立臺東大學教授
通訊作者: 鄭承昌，email: jengcc@gmail.com