Article

Educational and Psychological
Measurement
2019, Vol. 79(5) 874–882
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0013164418817803
journals.sagepub.com/home/epm
SAGE

Evaluation of Variance Inflation Factors in Regression Models Using Latent Variable Modeling Methods

Katerina M. Marcoulides[1] and Tenko Raykov[2]

Abstract

A procedure that can be used to evaluate the variance inflation factors and tolerance indices in linear regression models is discussed. The method permits both point and interval estimation of these factors and indices associated with explanatory variables considered for inclusion in a regression model. The approach makes use of popular latent variable modeling software to obtain these point and interval estimates. The procedure allows more informed evaluation of these quantities when addressing multicollinearity-related issues in empirical research using regression models. The method is illustrated on an empirical example using the popular software M*plus*. Results of a simulation study investigating the capabilities of the procedure are also presented.

Keywords

latent variable modeling, multicollinearity, regression model, simulation study, tolerance index, variance inflation factor

Regression modeling is extremely popular with researchers in the social, behavioral, clinical, educational, economic, business, marketing, organizational, and communication sciences (e.g., Raykov & Marcoulides, 2012). To appropriately examine the

[1] University of Florida, Gainesville, FL, USA
[2] Michigan State University, East Lansing, MI, USA

Corresponding Author:
Katerina M. Marcoulides, University of Florida, Research and Evaluation Methodology, 1215 Norman Hall, PO Box 117049, Gainesville, FL 32611, USA.
Email: k.marcoulides@coe.ufl.edu

unique explanatory power of a set of studied independent variables, researchers must pay careful attention and resolve the issue of multicollinearity (or near multicollinearity) as it can contribute markedly to inflated standard errors and spurious lack of statistically significant findings (e.g., Draper & Smith, 2012). Variance inflation factors (VIFs), or the closely related tolerance indices (TIs), are two relevant and frequently used quantities that may be consulted to examine individual predictors for potentially strong contributions to (near) multicollinearity (e.g., Wooldridge, 2015). Both VIFs and TIs are considered important quantities because they reflect estimates of the degree of interrelationship of an independent variable with other explanatory variables in a regression model (O'Brien, 2007).

An important limitation of currently widely used point estimates of VIFs and TIs that are readily available in statistical software programs (e.g., SAS, SPSS, Stata) is that they do not reflect the associated degree of instability of their estimation. Hence, when using what may at times be considered rough ''rules-of-thumb'' for VIFs (or TIs; e.g., Chatterjee & Simonoff, 2013), a researcher may miss a relevant finding for a predictor with a VIF that has been estimated as below a ''threshold'' assumed of importance while being higher in the studied population. This could for instance be due to sample size that may not be sufficiently large to afford more precise VIF (or TI) estimation and handling of the associated sampling error. In such cases, examination of interval estimates of the VIF and TI quantities may be very helpful as they may signal a potential problem with a predictor possibly contributing considerably to (near) multicollinearity.

Unfortunately, to date no readily applicable interval estimation procedures for VIFs or TIs have been made widely available for applied researchers. It is the aim of this article to discuss a directly utilizable method for this purpose. The approach is readily employed with the popular latent variable modeling software M*plus* (Muthén & Muthén, 2018) and can be used on a routine basis in empirical research using linear regression models.

Variance Inflation Factors and Tolerance Indices

The standard (multiple) linear regression analysis model (e.g., Draper & Smith, 2012) assumes that a response variable (i.e., dependent variable, outcome variable, or regressand), denoted Y, is related as follows to a given set of predictors (i.e., covariates, independent variables, explanatory variables, or regressors), designated $X_1, \ldots, X_k$:

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_kX_k + e, \qquad (1)$$

where the model parameters are the intercept a, the partial regression coefficients $b_1, \ldots, b_k$, and the variance of the residual term e, symbolized σ^2, that is, $\sigma^2 = Var(e)$, where $Var(.)$ denotes variance ($k \geq 1$). The associated and well-known standard assumptions also include homoscedasticity and exogeneity (e.g., Wooldridge, 2015).

When some of the predictors are involved in considerable linear relationships among themselves, standard errors for one or more individual partial regression coefficients can be unduly inflated. This tends to produce findings of possible lack of unique significance for substantively important regressors, in the context of other explanatory variables (e.g., Draper & Smith, 2012). As a means of gauging potential (near) multicollinearity, the variance inflation factors (VIFs)

$$V_j = 1/\left(1 - R_j^2\right) \tag{2}$$

have been proposed, where R_j^2 denotes the R^2 index when the jth explanatory variable is regressed on the remaining independent variables, or alternatively the use of the related tolerance indices (TIs)

$$T_j = 1/V_j \tag{3}$$

($j = 1, \ldots, k$; cf. Wooldridge, 2015). Informal, rough ''rules-of-thumb'' suggest that a predictor, X_j, with $V_j > 10$ or $T_j < 0.10$ values, may well be a cause of serious (near) multicollinearity. Other informal threshold criteria have also been suggested, endorsing that predictors with values above a VIF > 5 or a TI < 0.20 could also be contributing considerably to multicollinearity and generally deserve close inspection (e.g., Chatterjee & Simonoff, 2013; O'Brien, 2007). For example, Menard (1995) noted, ''A tolerance of less than 0.20 is cause for concern; a tolerance of less than 0.10 almost certainly indicates a serious collinearity problem'' (p. 66). Given that VIF is the inverse of TI, a tolerance value of 0.20 corresponds to what may be called the ''rule of 5'' and a tolerance value of 0.10 to the ''rule of 10'' with respect to the VIF.

If when fitted to a given data set the regression model (1) is found plausible (e.g., Raykov & Marcoulides, 2012), point estimates of the VIFs and TIs are readily obtained by substitution of the resulting R^2 indices into Equations (2) and (3), leading to ($j = 1, \ldots, k$)

$$\hat{V}_j = \frac{1}{1 - \hat{R}_j^2} \tag{4}$$

and

$$\hat{T}_j = 1 - \hat{R}_j^{\,2}, \tag{5}$$

where a circumflex is used to denote an estimate. These point estimates are straightforwardly obtained with widely circulated statistical software, such as SPSS, Stata, or SAS. Unfortunately, however, these and other statistical software do not provide the interval estimates associated with the VIFs and TIs. Yet their interval estimates can play an important role in empirical research. Specifically, confidence intervals (CIs) at prespecified confidence levels can reveal that certain predictors deserve more attention as potentially ‘‘offending’’ regressors than may be warranted by just examining their VIF or TI point estimates, $\hat{V}_j$ or $\hat{T}_j$, respectively ($1 \le j \le k$). This may in particular be the case when the upper endpoints of these intervals are above certain ‘‘thresholds,’’ such as say 5 or 10 mentioned above (e.g., Chatterjee & Simonoff, 2013). Such findings may be of relevance in studies with less than (fairly) large samples for predictors with say VIF point estimates close to but under a VIF ‘‘threshold’’ (e.g., 5). In these cases, the associated estimation imprecision or instability in the VIF point estimate (4) may cause a researcher to miss paying attention to a potentially relevant predictor, viz. one with a possibly true (population) VIF above that ‘‘threshold.’’ Examination then (and not only) of the confidence interval for the pertinent VIF index, say at an appropriate confidence level, can help sense that potential problem and thus contribute to a more informed decision and action on part of the researcher.

Point and Interval Estimation of Variance Inflation Factors and Tolerance Indices With Latent Variable Modeling Software

VIF and TI evaluation can be readily accomplished using the latent variable modeling (LVM; B. O. Muthén, 2002) framework, and in particular the popular LVM software *M*plus (Muthén & Muthén, 2018). To this end, while fitting the regression model of relevance, one can request the standardized solution that will provide the associated R^2 index along with a standard error. Employing on these R^2 and standard error values the initial monotone transformation approach given in Raykov and Marcoulides (2011), will render a corresponding large-sample CI for the R^2 index, denoted say (r_{lo}^2, r_{up}^2), where r_{lo}^2 and r_{up}^2 symbolize the lower and upper endpoint of the determined CI, respectively.

Next, based on Equation (2) and due to the fact that its right-hand side is a monotone increasing function of the R^2 index, the sought CI at the same confidence level for the VIF results by a corresponding inversion of the lower and upper endpoints of the above CI as

$$\left(1/\left(1-r_{lo}^2\right),\ 1/\left(1-r_{up}^2\right)\right). \tag{7}$$

The same level CI for the tolerance index, based on Equation (3) and the fact that this index is a decreasing function of the VIF, follows similarly as

$$\left(1-r_{up}^2, 1-r_{lo}^2\right). \tag{8}$$

The discussed CI construction procedure for the VIF and TI is readily conducted using the *R*-function ‘‘ci.vif_ti’’ provided in Appendix B. We note in passing that since the aforementioned monotone transformation approach yields upper and lower limits of the CI for the R^2 index that are within the interval (0, 1), the corresponding endpoints of the CIs in (7) and (8) will also reside within the same interval and could not fall outside it.

We demonstrate the application of the discussed VIF and TI evaluation procedure on empirical data in the next section and present subsequently the results of a simulation study on its performance.

Illustration on Empirical Data and a Simulation Study

A Numerical Data Example

To demonstrate the utility of the outlined method, we use an adapted data set from an ecological study concerned with examining daily average temperature in medium to large cities in the United States, denoted Y. We are interested in predicting this temperature using five relevant predictors, designated X_1 through X_5, which represent the number of manufacturing jobs, the number of inhabitants, the number of windy days per year, precipitation, and the number of sunny days per year, respectively (e.g., Hamilton, 2013). The analyzed data are available from the authors upon request.

Fitting first the pertinent multiple regression model,

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_5X_5 + e, \qquad (9)$$

and requesting the standardized solution, we obtain an R^2 estimate of 0.735 with a standard error of 0.037. (See Appendix A for the source code needed to accomplish these activities.)

Applying then on this estimate and standard error the R-function ‘‘ci.vif_ti’’ provided in Appendix B, we obtain the following point and interval estimates correspondingly for this model’s VIF and TI indices (95% CI follows point estimate):

$$\text{VIF} : 3.774(2.911, 5.025) \quad \text{and} \quad \text{TI} : 0.265(0.199, 0.343). \qquad (10)$$

These results suggest a potentially important finding for a researcher conducting this study. Specifically, the fact that the upper endpoint of the VIF CI is above 5 (while lower than 10; see earlier discussion) suggests that although there is no strong evidence for near multicollinearity it would be recommended to examine the used predictors for a possible relationship nearing linearity. Since city size is markedly correlated with number of manufacturing jobs and inhabitants, it may be suggested to consider possibly removing one of these two predictors from the model. When removing the ‘‘number of inhabitants’’ variable, for instance, the adjusted R^2 index (e.g., Raykov & Marcoulides, 2012) drops by .026 units down from .735, which may be viewed as not necessarily considerable. Although the drop in R^2 then to .706 is found to be significant, this result may be in part due to a sample size effect. Ultimately whether this predictor removal is to be carried out indeed is a decision that needs to be taken after thorough examination of its effect on the explanatory power of the model as well as the specifics of the research question and purpose of regression analysis conducted, and is best left to experts in the subject-matter domain.

Simulation Study and Results

To systematically evaluate the capability of the proposed procedure to supply informative point and interval VIF and TI index estimates, simulated data using Monte Carlo techniques were analyzed under a variety of conditions. All simulated data were generated using the Monte Carlo features available in M*plus* (L. K. Muthén & Muthén, 2018). We considered different magnitudes of relationships between predictor variables to represent different levels of collinearity. The simulated data were modeled after a study that examined the relationships between various patient health–related predictors (e.g., age, weight, triglycerides, high-density lipoprotein, and low-density lipoprotein levels) and measures of heart disease levels obtained from a sample of n = 95 patients at a California hospital (SAS Institute, 2014; data set ‘‘lipids.xls’’). In the original study, the magnitude of the correlation coefficients among the predictor variables ranged from 0.2 to 0.96. Intuitively, the changing amounts of collinearity should be reflected in the point and interval VIF and TI index estimates associated with the data.

Based on a review of the literature on similar past simulation studies, a number of factors were selected to be fixed across all data design conditions while other factors were varied (e.g., Vatcheva, Lee, McCormick, & Rahbar, 2016). Parameters that were fixed in the present simulations included the sample size, the magnitudes of the correlations between the predictors and the outcome variable, and the means and variances of the variables. Parameters that were varied in the simulations included the magnitude of the correlation between the predictors. Based on the recommendations of Vatcheva et al. (2016), these values were selected to reflect different degrees of multicollinearity that are commonly encountered in empirical research studies. Table 1 provides the different correlation matrices that were used in this simulation study to represent the different degrees of multicollinearity in the generated data along with the average obtained VIF and TI point and interval values across replications.

An examination of the various point and interval estimates displayed in Table 1 indicates several noteworthy findings. First, it is evident that the proposed procedure for generating interval estimates is able to provide informative supplementary insights over and above those furnished by only examining point estimates. For example, in the case of Model D, although the obtained VIF point estimate (VIF = 3.273) does not seem to suggests that there is evidence of multicollinearity following the earlier mentioned ‘‘rule of thumb’’ for VIF $>$ 5 (Chatterjee & Simonoff, 2013), the fact that the upper endpoint of the VIF CI [2.086, 5.760] is above 5 suggests (1) that there is evidence for potential near multicollinearity and (2) that it would be recommendable to examine the used predictors for a possible relationship nearing linearity. An examination of the correlation matrix used to simulate data for Model D reveals that there is indeed a sizeable correlation between variables X_1 and X_2. Model G is also another example of a situation where the obtained VIF point estimate (VIF = 4.1667) does not seem to reveal that there is evidence of multicollinearity. An examination of the upper endpoint of the VIF CI [2.081, 10.274] that is however above 10, indicates (1) that there is evidence for near multicollinearity and (2) that it would be recommendable in particular to examine the used predictors for a

Table 1. Correlations With Different Degrees of Multicollinearity Used for the Simulated Data (*n* = 100).

Model	Matrix specification	VIF CI	TI CI
A	$\begin{bmatrix} 1 & .5 & .1 & .2 \\ .5 & 1 & .1 & .2 \\ .1 & .1 & 1 & .4 \\ .2 & .2 & .4 & 1 \end{bmatrix}$	1.242 [1.089, 1.657]	.805 [.604, .918]
B	$\begin{bmatrix} 1 & .85 & .1 & .2 \\ .85 & 1 & .5 & .2 \\ .1 & .1 & 1 & .4 \\ .2 & .2 & .4 & 1 \end{bmatrix}$	1.532 [1.2664, 2.063]	.653 [.485, .790]
C	$\begin{bmatrix} 1 & .95 & .1 & .2 \\ .95 & 1 & .5 & .2 \\ .25 & .5 & 1 & .4 \\ .2 & .2 & .4 & 1 \end{bmatrix}$	1.974 [1.300, 4.156]	.506 [.241, .769]
D	$\begin{bmatrix} 1 & .85 & .5 & .2 \\ .85 & 1 & .85 & .2 \\ .5 & .85 & 1 & .4 \\ .2 & .2 & .4 & 1 \end{bmatrix}$	3.273 [2.086, 5.760]	.306 [.174, .479]
E	$\begin{bmatrix} 1 & .95 & .85 & .2 \\ .95 & 1 & .85 & .4 \\ .85 & .85 & 1 & .6 \\ .2 & .4 & .6 & 1 \end{bmatrix}$	15.303 [17.974, 257.105]	.065 [.004, .056]
F	$\begin{bmatrix} 1 & .5 & .1 & .2 \\ .5 & 1 & .1 & .4 \\ .1 & .1 & 1 & .6 \\ .2 & .4 & .6 & 1 \end{bmatrix}$	1.915 [1.451, 2.855]	.522 [.350, .689]
G	$\begin{bmatrix} 1 & .95 & .5 & .2 \\ .95 & 1 & .85 & .4 \\ .5 & .5 & 1 & .6 \\ .2 & .4 & .6 & 1 \end{bmatrix}$	4.167 [2.081, 10.274]	.240 [.097, .480]

Note. VIF = variance inflation factor; TI = tolerance index; CI = confidence interval.

possible relationship nearing linearity. In both of these cases, we may stress that through an examination of the interval estimates of the VIF (and the corresponding TI) quantities would it become evident that there is a potential problem with a predictor possibly contributing notably to (near) multicollinearity.

Conclusion

This article discussed a procedure for interval estimation of VIFs and TIs in regression models. The method is readily and widely applied using popular LVM software, such as M*plus*. The approach permits more informed decisions about potentially important near multicollinearity situations to be sensed in empirical research. A particularly useful feature of the procedure lies in the fact that it provides CIs at any prespecified confidence level for the VIFs and TIs. The method is applicable under the same assumptions made when using regression analysis (e.g., Wooldridge, 2015) and is best used with larger samples so that the underlying delta method and related standard error procedure applications are more trustworthy. We encourage future research with more comprehensive simulation studies that could shed additional light on the performance of the discussed interval estimation procedure for VIFs and TIs.

In conclusion, this article provides behavioral, social, educational, marketing, clinical, and organizational scholars with a procedure for evaluation of estimation instability of routinely output multicollinearity-related indices by statistical software.

Appendix A

Mplus Source Code for Evaluation of Variance Inflation Factors and Tolerance Indices in Regression Models

```
TITLE:       FITTING A REGRESSION MODEL, FOR INTERVAL ESTIMATION OF
             R-SQUARE.
DATA:        FILE = < NAME OF RAW DATA FILE > ;
VARIABLE:    NAMES = ID SO2 TEMP MANUF POP WIND PRECIP DAYS;
             USEV = TEMP-DAYS;
ANALYSIS:    ESTIMATOR = MLR;
MODEL:       TEMP ON MANUF-DAYS;
OUTPUT:      STANDARDIZED;
```

Appendix B

R-*Function for Interval Estimation of Variance Inflation Factors and Tolerance Indices*

```
ci.vif_ti = function(r, se){ # R-function for CI construction for VIF and TI.
  l = log(r/(1-r)) # see Raykov & Marcoulides (2011, ch. 7) for details.
  sel = se/(r*(1-r))
  ci_l_lo = l-1.96*sel
  ci_l_up = l+1.96*sel
  ci_lo = 1/(1+exp(-ci_l_lo))
  ci_up = 1/(1+exp(-ci_l_up))
  vif = 1/(1-r)
  ti = 1-r
  vif_ti_pe_ci = c(vif, 1/(1-ci_lo), 1/(1-ci_up), ti, 1-ci_up, 1-ci_lo)
  vif_ti_pe_ci # prints to screen VIF, its 95%-CI, TI, and its 95%-CI, in this
  order.
}
```

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Chatterjee, S., & Simonoff, J. S. (2013). *Handbook of regression analysis*. New York, NY: Wiley.

Draper, N. R., & Smith, H. (2012). *Applied regression analysis*. New York, NY: Wiley.

Hamilton, L. C. (2013). *Statistics with Stata*. Boston, MA: Brooks/Cole.

Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*, 87-117.

Muthén, L. K., & Muthén, B. O. (2018). M*plus user's guide*. Los Angeles, CA: Muthén & Muthén.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*, 673-690.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.

Raykov, T., & Marcoulides, G. A. (2012). *Basic statistics: An introduction with R*. New York, NY: Rowman & Littlefield.

SAS Institute, Inc. (2014). *SAS/STAT® 13.2 user's guide*. Cary, NC: Author.

Vatcheva, K. P., Lee, M., McCormick, J. B., & Rahbar, M. H. (2016). Multicollinearity in regression analysis conducted in epidemiologic studies. *Epidemiology*, *6*, 227-247.

Wooldridge, J. M. (2015). *Introductory econometrics. A modern approach*. Boston, MA: Cengage Learning.