


Collinearity diagnostic applied in ridge estimation through the variance inflation factor

Roman Salmerón Gómez^a, José García Pérez^b, María Del Mar López Martín^c and Catalina García García^a 

^aDepartment of Quantitative Methods for Economic and Business, Granada University, Granada, Spain;

^bDepartment of Economic and Business, Almería University, Almería, Spain; ^cDepartment of Didactics of Mathematics, Granada University, Granada, Spain

ABSTRACT

The variance inflation factor (VIF) is used to detect the presence of linear relationships between two or more independent variables (i.e. collinearity) in the multiple linear regression model. However, the traditionally used VIF definitions encounter some problems when extended to the case of the ridge estimation (RE). This paper presents an extension of the VIF in RE by providing two alternative VIF expressions that overcome these problems in the general case. Some characteristics of these expressions are also presented and compared with the traditional expression. The results are illustrated with an economic example in the case of three independent variables and with a Monte Carlo simulation for the general case.

ARTICLE HISTORY

Received 22 April 2015

Accepted 12 November 2015

KEYWORDS

Multiple linear regression; collinearity; ridge regression; multivariables

1. Introduction

The collinearity problem is the existence of linear relationships between two or more independent variables in model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad (1)$$

where $E[\mathbf{u}] = \mathbf{0}$, $E[\mathbf{u}\mathbf{u}'] = \sigma^2\mathbf{I}$. It is considered that the variables in Equation (1) are standardized, therefore $\mathbf{X}'\mathbf{X}$ is the correlation matrix and $\mathbf{X}'\mathbf{Y}$ is the vector of correlation coefficients of the response variable with each of the explanatory variables. Collinearity will be perfect or approximate depending on the kind of relation. In the first case, the model does not satisfy the full range condition and has infinite solutions, while, in the second one, although the condition is fulfilled, the estimation will be unstable and the variance of the estimators may be large compared to the values of the estimated parameters that can be insignificant or have the wrong sing. Obviously, the second case is troubling. Remember that the collinearity is a data problem belonging to numerical analysis research area that can be found in any field.

The variance inflation factor (VIF) [8,23,32] has been widely applied in scientific literature to diagnose the existence of collinearity, although it is possible to find other measures such as the eigenvalues through the Condition Index [1,3], the condition number (CN)

[2,3], the variance decomposition proportions to analyze the correlations between different vectors and their angles [29,34], the red indicator [20], the corrected VIF [5], or the use of biplot method in the visualization diagnostic of multicollinearity problems called collinearity biplot [10].

Focusing on the VIF, the definition provided by Theil [32] allows to measure the impact of collinearity of the variable \mathbf{X}_i , $i = 1, \dots, p$, with the rest of the independent variables

$$\text{VIF}(i) = \frac{1}{1 - R_i^2}, \quad i = 1, \dots, p, \quad (2)$$

where R_i^2 is the coefficient of determination of \mathbf{X}_i on the rest of independent variables. Marquardt [23] defined the VIF as the elements of the principal diagonal of the inverse correlation matrix. Thus, the VIF will be the elements of the diagonal of the inverse of $\mathbf{X}'\mathbf{X}$ since the correlation matrix of the independent variables is the matrix $\mathbf{X}'\mathbf{X}$ when the data are standardized. This measure has well-known weaknesses that should be noted. Firstly, the controversy about the use of centered or not centered data that has been recently revised by García *et al.* [13]. Secondly, the fact that there is no measure to know how closely R_i^2 must be to 1 to imply collinearity [2]. Furthermore, the VIF is not resistant to the present of high leverage points (outliers). Finally, it is generally accepted that values of VIF higher than 10 indicate severe collinearity [19] but this rule of thumb lacks a theoretical basis. Indeed, taking into account the expression of the estimated variance of the estimated parameters

$$\widehat{\text{var}}(\hat{\beta}_i) = \frac{\hat{\sigma}_u^2}{n \text{var}(X_i)} \cdot \text{VIF}(i), \quad i = 1, \dots, p, \quad (3)$$

where n is the number of observations, high values of VIF could not imply high estimated variance since it can be countered by the ratio of the variance of the error terms divided by the variation in the respective independent variable. Thus, the variance of the error is also an important factor to get high variance and, for this reason, practitioners should apply statistical skills for model modification to minimize it and thereby the variance of the estimated parameters.

Focusing on the diagnose, and not in the solution of the collinearity, the VIF is widely applied as can be noted from the paper of O'Brien [27] with more than 1700 references in a great variety of fields. The concept of VIF was generalized in ordinary least squares (OLS) by Fox and Monette [8], who defined the generalized variance inflator factor (GVIF) as the measure of the impact of collinearity on the square of the length of the joint confidence region (two or more coefficients)

$$\text{GVIF} = \frac{|\mathbf{R}_1||\mathbf{R}_2|}{|\mathbf{R}|}, \quad (4)$$

where $|\cdot|$ is the matrix determinant, \mathbf{R}_1 is the correlation matrix of a particular set of regressors, \mathbf{X}_1 , \mathbf{R}_2 is the correlation matrix of the rest of the regressors, \mathbf{X}_2 , and \mathbf{R} is the correlation matrix of all the regressors. Note that when the number of variables in \mathbf{X}_1 is equal to one, the initial expression is simplified and the GVIF coincides with the VIF (see, for example, [4]).

On the other hand, the ridge estimation (RE) is a widespread method to overcome the problem of collinearity defining a class of estimators depending on the non-negative scalar

parameter λ

$$\hat{\beta}(\lambda) = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}. \quad (5)$$

Its covariance matrix is

$$\text{var}(\hat{\beta}(\lambda)) = \sigma^2(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}. \quad (6)$$

The estimator given in Equation (5) is a biased estimator when $\lambda > 0$ and when $\lambda = 0$ coincides with the OLS estimator. Despite earlier usage in numerical analysis [22,30], the ridge estimator is an interesting topic usually labeled in statistic and econometric and with applications in many different fields such as medicine, physics and chemistry. Indeed, McDonald [24] conducted a detailed study of the scientific literature on the ridge estimator since its presentation in the seminal papers of Hoerl and Kennard [16,17] and concluded that more than 240 works have been published (see [24]) in prestigious journals.

Once the parameter λ is selected and RE is applied, it is necessary to calculate again the value of a diagnostic measure to check if collinearity has been mitigated enough. This fact justifies the extension of the collinearity indicators to be applied after RE. García *et al.* [12] extended the VIF for the case $p = 2$ and the condition number was extended by García *et al.* [14]. The purpose of this article is to show the deficiencies obtained when applying in RE the definitions of VIF originally created to OLS estimation and propose an alternative expression for $p > 2$ that verifies some desirable properties.

The structure is as follows: Section 2 presents the problems when applying the traditional VIF definitions in RE by using an example. In Section 3 the VIF of surrogate ridge model and the VIF expression from the vector that generate the matrix of the ridge estimators for $p = 3$ are calculated. Some properties of both VIFs are also shown and is calculated as an explicit expression based on the correlation coefficients for the GVIF when $p = 3$ in RE. These expressions are used to analyze the presence of multicollinearity in real economic data in Section 4. The results are compared with the extension of the condition number in RE. Due to the difficulty of obtaining expressions for these VIFs for $p > 3$, Section 5 shows how to study the presence of multicollinearity in a model with $p > 3$ obtaining the above expressions computationally for any value of p . It also highlights some limitations of using generalized VIF in RE. Finally, Section 6 resumes the main contributions of the work.

2. VIF extension to RE

The first extension of the VIF associated with the ridge estimator was given by Marquardt [23] who proposed detecting the presence of collinearity by using the diagonal elements of the matrix $(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}$ as the VIF. Note that Marquardt automatically extended the definition of the VIF in OLS regression to RE. Kutner *et al.* [21] proposed a definition of the VIF in RE that coincides with the extension of the Marquardt's VIF definition in RE (see Appendix 1). However, the expression obtained for the VIF(λ) (see [23] for the case of two variables) does not satisfy the condition of being larger than one (see [11,13,31]). This is the first problem presented by the expression proposed by Marquardt. Furthermore, the expression given by Theil [32] cannot be calculated since initially we do not know the matrix \mathbf{Z} from which to obtain $\mathbf{Z}'\mathbf{Z} = \mathbf{X}'\mathbf{X} + \lambda\mathbf{I}$. The same occurs to the one proposed by Fox and Monette [8] although in this case Friendly [9] provides the correlation

Table 1. Values of VIFs of Marquardt ($\text{REMVIF}(\lambda, i)$), Fox and Monette ($\text{REGVIF}(\lambda, i)$) and Theil ($\text{VIF}(i)$) of the i th variable for $\lambda = 0$ and $\lambda = 0.1$.

| | $\text{REMVIF}(0, i)$ | $\text{REGVIF}(0, i)$ | $\text{VIF}(i)$ | $\text{REMVIF}(0.1, i)$ | $\text{REGVIF}(0.1, i)$ |
|---------|-----------------------|-----------------------|-----------------|-------------------------|-------------------------|
| $i = 1$ | 154.9487 | 154.9487 | 154.9487 | 0.6272 | 1.724408 |
| $i = 2$ | 37.2681 | 37.2681 | 37.2681 | 1.0804 | 1.699081 |
| $i = 3$ | 222.8143 | 222.8143 | 222.8143 | 0.3658 | 1.254078 |

matrix of the data by conveniently transforming the covariance matrix (6). By using the R statistical environment, the *genridge* package (see [9]) allows us to calculate the GVIF in RE by using the following expression:

$$\text{GVIF}(i) = \frac{|\mathbf{R}_{-i}|}{|\mathbf{R}|}, \quad i = 1, \dots, p, \quad (7)$$

where \mathbf{R} denotes the correlation matrix among all the columns of \mathbf{X} and \mathbf{R}_{-i} is the resulting matrix by eliminating the i th row and i th column in matrix \mathbf{R} . Remember that $|\mathbf{R}| = 1$ for orthogonal data and $|\mathbf{R}| = 0$ for perfectly collinear data (see [28]).

To illustrate all these affirmations, we will use the model previously used by Wissel [35] relative to credit in American people with the following variables: the **total** mortgage dept outstanding, \mathbf{Y} , personal consumption, \mathbf{X}_1 , personal incomes, \mathbf{X}_2 , and consumer credit outstanding, \mathbf{X}_3 . All variables are measured in billions of dollars. We have enlarged the sample by using data from 1995 to 2011 obtained from Economic Reports of the President [6]. The values shown in Table 1 are the extension to RE of the VIF definition given by Marquardt [23], denoted as REMVIF, and the extension of the general VIF definition given by Fox and Monette [8], denoted as REGVIF.

Note that all the definitions presented lead to the same result when the OLS estimation is applied ($\lambda = 0$). However, when these definitions are extended and applied in RE ($\lambda > 0$) they provide different results. The source of the problem may be that when we use the OLS estimator, the matrix of independent variables, \mathbf{X} , and the correlation matrix of the regressors, \mathbf{R} , are known. This situation is very different when we work with the ridge estimator: the data which generate the matrix $\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}$ are initially unknown and consequently we do not have the information to obtain the determination coefficient to calculate the corresponding VIFs (expression proposed by Theil [32]) or the correlation matrix of the data (expression proposed by Fox and Monette [8]). These limitations may be the cause to that numerous authors have followed the proposal by Marquardt [23] to extend the concept of VIF to RE. This fact leads not only to values of VIF lesser than one but also they do not have the desirable property of being monotonic. Furthermore, as shown in the example, every measure to diagnostic collinearity leads to a different solution and, in some cases, to a different conclusion. In the next section, we present an alternative expression for the VIF in RE with $p > 2$ and analyze its properties. We will use this same example to illustrate empirically the application of the proposed methodology.

3. RESVIF, REVIF and REGVIF in models with three exogenous variables

As mentioned above, the main problem in extending the definitions used in the OLS estimation to calculate the VIF in RE is that the matrix \mathbf{Z} that verifies $\mathbf{Z}'\mathbf{Z} = \mathbf{X}'\mathbf{X} + \lambda\mathbf{I}$ is

initially unknown. This question was initially solved by García *et al.* [12] for the case $p = 2$. In this work we present the extension to the general case beginning for the case $p = 3$.

This problem can be solved from the matrix \mathbf{X}_S , verifying $\mathbf{X}'_S \mathbf{X}_S = \mathbf{X}' \mathbf{X} + \lambda \mathbf{I}$, which was provided by Jensen and Ramirez [18] in the presentation of the surrogate RE. Note that the matrix \mathbf{X}_S keeps the dimension of the matrix \mathbf{X} and it is orthonormalized. The VIF obtained from \mathbf{X}_S will be named as RESVIF.

Within the ridge estimator methodology, we can use the augmented model proposal of Marquardt [23] to obtain by OLS the ridge estimator and use the matrix $\mathbf{X}_Z = \left(\frac{\mathbf{X}}{\sqrt{\lambda \mathbf{I}}} \right)$ instead of the matrix \mathbf{X}_S . We will call the VIF calculated from this matrix REVIF.

In Appendix 2 we obtain explicit expressions for the RESVIF and REVIF when $p = 3$. Both expressions are obtained from the coefficient of determination of an auxiliary regression. Some properties of both VIFs are also shown. To finish, we obtain an explicit expression based on the correlation coefficients for REGVIF when $p = 3$. Follow the original expressions to calculate the RESVIF, REVIF and REGVIF are presented as a major contribution of this paper:

- Ridge estimator surrogate VIF (RESVIF) for the i th variable is

$$\text{RESVIF}(\lambda, i) = \frac{(1 + \lambda)[(1 + \lambda)^2 - \rho_{jk}^2]}{(1 + \lambda)[(1 + \lambda)^2 - \rho_{ij}^2 - \rho_{ik}^2 - \rho_{jk}^2] + 2\rho_{ij}\rho_{jk}\rho_{ik}}. \quad (8)$$

It is verified that $\text{RESVIF}(\lambda, i) \geq 1$, $\lim_{\lambda \rightarrow \infty} \text{RESVIF}(\lambda, i) = 1$ and the RESVIF is monotone decreasing with increasing λ .

- Ridge estimator VIF (REVIF) for the i th variable is

$$\begin{aligned} &\text{REVIF}(\lambda, i) \\ &= \frac{n + 3 + \lambda(n + 2)}{(n + 3) \left[\frac{\lambda(2\rho_{ij} + 2\rho_{ik} - (\lambda + \rho_{jk} - n)) + (n + 3)(\rho_{jk} + 1)}{B} + \frac{2\rho_{ij}\rho_{ik}D - (\rho_{ij}^2 + \rho_{ik}^2)C}{A} \right] + \lambda(n + 2)}. \end{aligned} \quad (9)$$

It is verified that $\text{REVIF}(\lambda, i) \geq 1$, $\lim_{\lambda \rightarrow \infty} \text{REVIF}(\lambda, i) = 1$ and the REVIF is monotone decreasing with increasing λ .

- Ridge estimator generalized VIF (REGVIF) for the i th variable is

$$\text{REGVIF}(i) = \frac{1 - \rho_{jk}^2}{1 - \rho_{jk}^2 - \rho_{ij}^2 - \rho_{ik}^2 + 2\rho_{jk}\rho_{ij}\rho_{ik}}. \quad (10)$$

Note that ρ_{ij} represents the correlation between the variables $\mathbf{X}(i)$ and $\mathbf{X}(j)$ with $i, j, k \in \{1, 2, 3\}$, $i \neq j$, $i \neq k$ and $j \neq k$.

It can be easily demonstrated for $\lambda = 0$ that

$$\text{RESVIF}(0, i) = \text{REVIF}(0, i) = \text{REGVIF}(i) = \frac{1 - \rho_{jk}^2}{1 - \rho_{jk}^2 - \rho_{ij}^2 - \rho_{ik}^2 + 2\rho_{jk}\rho_{ij}\rho_{ik}} = \text{VIF}(i). \quad (11)$$

4. Obtaining the VIF in economic data

In Section 2, we showed the deficiencies obtained when applying in RE the definitions of VIF originally created to OLS estimation by using the model relative to credit in American people previously applied by Wissel [35] but enlarging the sample with data from 1995 to 2011 obtained from Economic Reports of the President [6]. Please find data in Appendix 4. Now we will use this same example to illustrate how to calculate the REVIF and compare it with the RESVIF and the REMVIF.

From this data, the multiple linear estimation of model (1) leads to the following results:

$$\hat{Y} = \underset{(8995)}{-920.9} - \underset{(2.399)}{1.396 \cdot X_1} + \underset{(0.6057)}{0.8886 \cdot X_2} + \underset{(0.00639)}{0.00667 \cdot X_3},$$

with $R^2 = 0.9583$ and $F_{\text{exp}} = 99.65$. Note that none parameter is statistically significant but they are not simultaneously zero (joint significance). We also observe that the coefficient of determination is very high. These results are typical of models with collinearity. As was noted in the introduction, a good specification of the model is a first relevant step to solve these problems. However, the collinearity may persist.

In this case, the correlation matrix shows very high correlations between all independent variables

$$R = \begin{pmatrix} 1 & 0.9795893 & 0.9966152 \\ 0.9795893 & 1 & 0.9858508 \\ 0.9966152 & 0.9858508 & 1 \end{pmatrix}. \quad (12)$$

This high correlation was expected since the independent variables (personal consumption, X_1 , personal incomes, X_2 , and consumer credit outstanding, X_3) seem to be very related from an economic point of view. Indeed, Wissel [35] recognized that the data were chosen to obtain multicollinearity. In addition, by calculating the R_i^2 , $i = 1, 2, 3$, we obtain the VIF of each variable which is always over 10 indicating the existence of collinearity (see Table 2).

According to the interpretation given by Fox [7] about the VIF, we can state, for example, that the confidence interval for β_1 and β_3 is $\sqrt{154.9487} = 12.4478$ and $\sqrt{222.8143} = 14.9269$ times greater, respectively, than if there were no multicollinearity. All this suggests the existence of collinearity in the proposed model which seems ideal to use the expression obtained in Section 3.

The VIFs of each variable, using the different expressions REMVIF, RESVIF, REVIF, REGVIF (from standardized variables) and REGVIF (from typified¹ variables) are represented in Figure 1 for $\lambda = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$.

Note that for λ varying between 0.1 and 1, the REMVIF of the variables X_1 and X_3 are always less than 1 while for variable X_2 only takes values higher than one when $\lambda = 0.1$. These results contradict the consequences obtained from expression (2). On the other

Table 2. Values of VIF.

| | $i = 1$ | $i = 2$ | $i = 3$ |
|--------------------------------|----------|---------|----------|
| R_i^2 | 0.9935 | 0.9731 | 0.9955 |
| $VIF(i) = \frac{1}{1 - R_i^2}$ | 154.9487 | 37.2681 | 222.8143 |

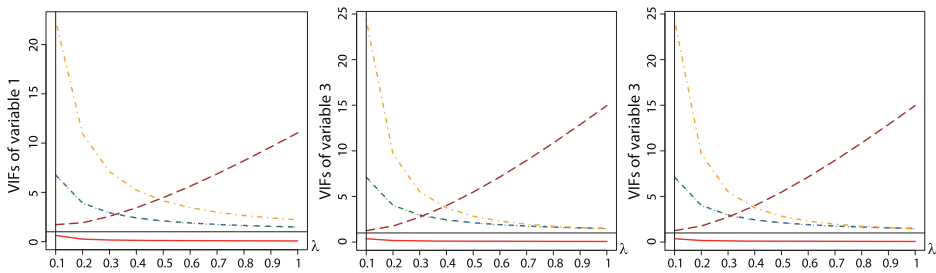


Figure 1. REMVIF (red-solid); RESVIF (blue-dashed); REVIF (green-dotted); REGVIF from standardized variables (brown-dotdash) and REGVIF from typified variables (orange-dotdash) join the line $y = 1$ [Colour online].

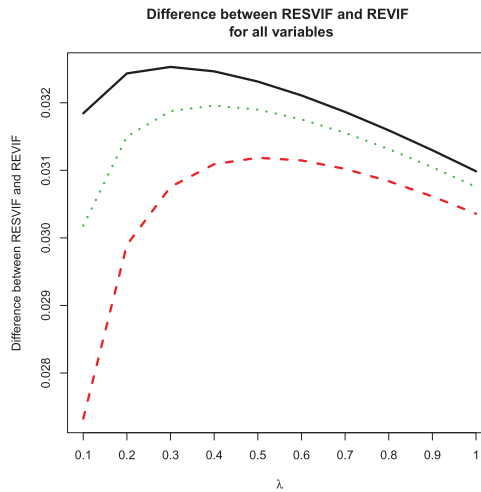


Figure 2. Representing the difference between RESVIF and REVIF for $\lambda = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$ for each variables [Colour online].

hand, the values of REGVIF are different from the one obtained for the RESVIF and REVIF. This difference is very significative when variables are standardized.

In Section A.2, expression (A18) shows that the RESVIF and the REVIF coincide when $n \rightarrow +\infty$. However, in Figure 1 we can see that both VIFs graphs overlap and in this case the value of n is only 17. Furthermore, Figure 2 shows the difference between RESVIF and REVIF for $\lambda = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$. Note that the differences are of the order of 10^{-2} . So a large sample size is not necessary to allow the property to be verified.

In Table 3 we study in more detail the results obtained when λ varies between 0 and 0.1. Note that from $\lambda = 0.02$, all values of REMVIF are less than 10 while this does not occur in RESVIF and REVIF until $\lambda = 0.07$ (see Tables 4 and 5). Following REMVIF this fact would imply that, for $\lambda = 0.02, 0.03, 0.04, 0.05, 0.06$, the problem of collinearity has been mitigated when in fact it has not. Note that when the parameter λ is equal to zero the values of VIFs are equal since in this situation the original model has been estimated by OLS. Table 6 presents the condition number (CN) extended from OLS to RE without further considerations and the $CN(\lambda)$ extended to RE taking into considerations the problems of this issue as shown by García *et al.* [12].

Table 3. Values of REMVIF for $\lambda = 0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1$.

| λ | X_1 | X_2 | X_3 |
|-----------|-------------|-----------|-------------|
| 0 | 154.9487572 | 37.268158 | 222.8143143 |
| 0.01 | 12.5436403 | 14.231998 | 11.5699764 |
| 0.02 | 5.6365410 | 8.271826 | 4.1167910 |
| 0.03 | 3.3836467 | 5.440672 | 2.1973748 |
| 0.04 | 2.3043321 | 3.865205 | 1.4041879 |
| 0.05 | 1.6909349 | 2.897617 | 0.9950503 |
| 0.06 | 1.3055132 | 2.260648 | 0.7546938 |
| 0.07 | 1.0463587 | 1.818998 | 0.6007831 |
| 0.08 | 0.8632570 | 1.500161 | 0.4959592 |
| 0.09 | 0.7288631 | 1.262416 | 0.4211669 |
| 0.1 | 0.6271743 | 1.080372 | 0.3658185 |

Table 4. Values of RESVIF for $\lambda = 0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1$.

| λ | X_1 | X_2 | X_3 |
|-----------|------------|-----------|------------|
| 0 | 154.948757 | 37.268158 | 222.814314 |
| 0.01 | 39.418727 | 21.965852 | 49.483674 |
| 0.02 | 24.137728 | 16.617822 | 28.474403 |
| 0.03 | 17.743912 | 13.489267 | 20.197535 |
| 0.04 | 14.162010 | 11.407017 | 15.750795 |
| 0.05 | 11.852657 | 9.915907 | 12.969567 |
| 0.06 | 10.233564 | 8.793964 | 11.063770 |
| 0.07 | 9.032913 | 7.918666 | 9.675491 |
| 0.08 | 8.105895 | 7.216529 | 8.618786 |
| 0.09 | 7.367971 | 6.640709 | 7.787377 |
| 0.1 | 6.766345 | 6.159898 | 7.116079 |

Table 5. Values of REVIF for $\lambda = 0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1$.

| λ | X_1 | X_2 | X_3 |
|-----------|------------|-----------|------------|
| 0 | 154.948757 | 37.268158 | 222.814314 |
| 0.01 | 39.399287 | 21.955079 | 49.459193 |
| 0.02 | 24.114188 | 16.601705 | 28.446546 |
| 0.03 | 17.718249 | 13.469864 | 20.168229 |
| 0.04 | 14.135007 | 11.385386 | 15.720666 |
| 0.05 | 11.824723 | 9.892666 | 12.938903 |
| 0.06 | 10.204944 | 8.769506 | 11.032729 |
| 0.07 | 9.003767 | 7.893256 | 9.644171 |
| 0.08 | 8.076331 | 7.190354 | 8.587250 |
| 0.09 | 7.338069 | 6.613907 | 7.755671 |
| 0.1 | 6.736164 | 6.132574 | 7.084235 |

From these results we can highlight the following comments:

- The concept of VIF in OLS is based on expression (2) and it has to be kept when extending to RE.
- The REMVIF presents values lesser than one unrespecting the concept of VIF. Furthermore, the REMVIF goes rapidly down the rules of thumb of VIF and it can lead to think that the multicollinearity is mitigated when it is not.

Table 6. Values of REVIF and CN for $\lambda = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$.

| λ | REVIF(1) | REVIF(2) | REVIF(3) | CN | CN(λ) |
|-----------|----------|----------|----------|---------|-----------------|
| 0 | 154.9488 | 37.2682 | 222.8143 | 75.1695 | 33.1943 |
| 0.1 | 6.7362 | 6.1326 | 7.0842 | 64.5731 | 18.6334 |
| 0.2 | 3.8902 | 3.713 | 3.9924 | 57.48 | 14.3649 |
| 0.3 | 2.8876 | 2.802 | 2.937 | 52.3065 | 12.1218 |
| 0.4 | 2.3768 | 2.3257 | 2.4063 | 48.3183 | 10.6853 |
| 0.5 | 2.0684 | 2.034 | 2.0882 | 45.1227 | 9.6652 |
| 0.6 | 1.8626 | 1.8377 | 1.8769 | 42.4876 | 8.8929 |
| 0.7 | 1.716 | 1.6971 | 1.7269 | 40.2664 | 8.2823 |
| 0.8 | 1.6067 | 1.5918 | 1.6153 | 38.3609 | 7.7838 |
| 0.9 | 1.5223 | 1.5101 | 1.5293 | 36.7028 | 7.3669 |
| 1 | 1.4553 | 1.4452 | 1.4611 | 35.2428 | 7.0117 |

- The REVIF coincides asymptotically with the RESVIF, and for values of n equal to 17, as in the numerical example, the difference is depreciable.
- The values obtained to the REGVIF (with original, typified or standardized data) do not coincide with the values obtained for the RESVIF or the REVIF.
- From Table 6 we see that the CN and CN(λ) are decreasing. While CN presents all values higher than 30, the CN(λ) takes values lesser than 10 from $\lambda > 0.4$. Thus, the CN(λ) and the REVIF lead to similar conclusions in the diagnostic of collinearity. However, the CN extended without further considerations indicates that the collinearity has not been mitigated even for $\lambda = 1$.

Finally, we would like to emphasize that in the case of the ridge estimator, the calculation of the VIFs proposed by Theil [32] (see expression (2)), Marquardt [23] (elements of the principal diagonal of the inverse correlation matrix) and Fox and Monette [8] (see expression (7)) leads us to obtain the same results obtained in the calculation of REVIF. The REVIF is monotonically decreasing in λ , is higher than 1 and it verifies that $\lim_{\lambda \rightarrow \infty} \text{REVIF}(\lambda, i) = 1$ and $\text{REVIF}(0, i) = \text{VIF}(i)$.

Then, we propose to use the REVIF instead of traditionally used REMVIF and REGVIF. The RESVIF, associated to surrogate ridge model, can be also recommended due to its simple calculation and its asymptotic equivalence with the REVIF but taking into consideration that it is associated to OLS surrogate ridge model.

5. A Monte Carlo simulation

In the last section we have presented an economic example with three exogenous variables and standardized data obtaining the REMVIF, RESVIF, REVIF and REGVIF with standardized data finding certain anomalies in the results obtained by the REGVIF since it increases when λ increases (see Figure 1), which does not make sense. Due to this fact, we calculate the REGVIF from the original data obtaining different results even for $\lambda = 0$. We also calculate the REGVIF from typified data and we have included this case in the work since the obtained results coincide with the results provided by the package *genridge* in R software. In this last case we obtain values of REGVIF different from the one obtained for RESVIF and REVIF. In this section, we will study in more detail this case and use a Monte Carlo simulation to present algorithms to obtain computationally the different VIFs for models with multiple independent variables.

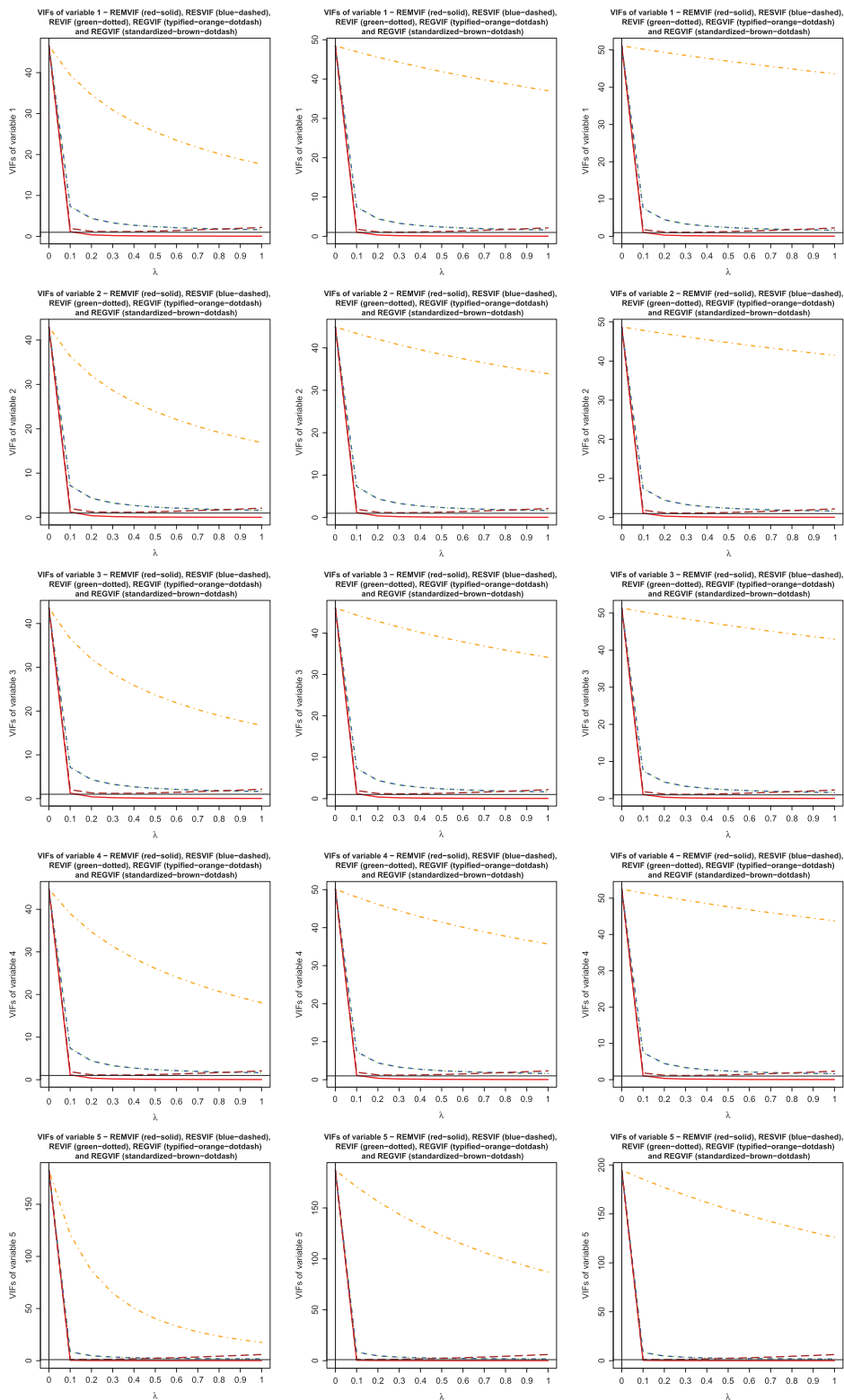


Figure 3. VIFs for all variables (first, second and third column corresponds to $n = 100, 500, 1000$, respectively) [Colour online].

Table 7. Difference between RESVIF and REVIF depending on sample size.

| Sample size | Minimum difference | Maximum difference |
|-------------|--------------------|--------------------|
| 100 | 0.0064 | 0.0074 |
| 500 | 0.0013 | 0.0015 |
| 1000 | 0.00068 | 0.00078 |

Following [15,26,33], we generate five independent variables using

$$\mathbf{X}(i) = \sqrt{1 - \gamma^2} \mathbf{Z}(i) + \gamma \mathbf{Z}(5), \quad i = 1, 2, 3, 4, 5, \quad (13)$$

where $\mathbf{Z}(i)$ are independent pseudo-random numbers distributed as $N(0, 100)$ and γ is specified so that the correlation between two any independent variable is given by γ^2 . The dependent variable is generated as

$$\mathbf{Y} = \mathbf{X}(1) + \mathbf{X}(2) + \mathbf{X}(3) + \mathbf{X}(4) + \mathbf{X}(5) + \mathbf{u}, \quad (14)$$

where \mathbf{u} are independent pseudo-random numbers distributed as $N(0, 1)$. Finally, we used three sample sizes: 100, 500 and 1000.

Since there are five independent variables it is necessary to use the algorithms proposed in Appendix 3. The results are shown in Figure 3 for $\lambda = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$. It is observed that:

- The REGVIF from standardized variables, as already mentioned, increases when λ increases.
- The REGVIF from typified variables increases when n increases.
- The REGVIF from typified variables is never lesser than the established limit 10. This means that the collinearity has not been mitigated.
- The other VIFs show the same behavior: (a) RESVIF and REVIF graphs overlap and (b) REMVIF takes values less than 1.
- The difference between RESVIF and REVIF decreases when n increases (see Table 7).

The first three points lead us to believe that the application of REGVIF (from typified or standardized data) is not appropriate in RE. This is not a minor issue since the package *genridge* in R software calculated the VIF in RE following this definition.

6. Conclusions

If all definitions of VIF [8,21,23,32] lead to the same results under OLS estimation, its extension to RE should also lead to the same result. However, we have obtained different results and conclusions up to the point of considering that the collinearity problem is solved when it still persists.

The problem is that both Marquardt [23] and Fox and Monette [8] use the covariance matrix of RE by extending its initial definition (in the case of Marquardt) or obtaining the correlation matrix (in the case of Fox and Monette). All extensions lead to the same results if instead of selecting this matrix as a starting point we choose the ridge regressors matrix (see expressions in [23]), which establishes the equivalence between RE and an alternative OLS

estimation. Thus, we can conclude that the expression proposed in this paper to calculate the VIF in RE is the right one. That is, the VIF associated with the ridge estimator should be calculated by the REVIF expression.

Furthermore, it has been shown that RESVIF (associated to ridge surrogate estimator) is easier to calculate than the REVIF and it can be its substitute since the differences are small, at least they are quoted to one hundredth.

When extending the $VIF(i)$ to ridge regression, the $VIF(\lambda, i)$ should be monotonically decreasing in λ , higher than one and it has to verify that $\lim_{\lambda \rightarrow \infty} REVIF(\lambda, i) = 1$ and $VIF(0, i) = VIF(i)$. These conditions are only verified by the REVIF (associated to ridge estimator) and the RESVIF (associated to surrogate ridge estimator). Although the RESVIF verifies these conditions and also presents desirable monotone properties, it is possible to get these same conditions within RE if the concept of VIF is appropriately extended. Thus, within RE the suitable extension should be the REVIF.

It was shown that the use of REGVIF in RE should be reviewed (a) from standardized variables major anomalies are obtained and (b) from typified variables (methodology used in R software) are obtained different values to those obtained by RESVIF and REVIF, and as shown in the simulation section can lead to erroneous conclusions.

To sum up, the main contribution of this paper is to present the expressions and algorithms necessary to diagnose correctly the collinearity through the VIF after the application of the RE for p independent variables tending to make life simpler for those who come across collinearity issues in their regression model.

Note

1. Note that a standardized variable is the value of the variable minus its mean, divided by the square root of the number of observations multiplied by its variance while a typified variable is the value of the variable minus its mean, divided by its standard deviation.

Acknowledgments

We would like to thank referees for their detailed, encouraging and constructive reviews of our paper which clearly contributed to improving both the structure and the content of this manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Catalina García García  <http://orcid.org/0000-0003-1622-3877>

References

- [1] D.A. Belsley, *Assessing the presence of harmful collinearity and other forms of weak data through a test for signal-to-noise*, J. Econ. 20 (1982), pp. 211–253.
- [2] D.A. Belsley, *Conditioning Diagnostics: Collinearity and Weak Data in Regression*, John Wiley, New York, 1991.
- [3] D.A. Belsley, E. Kuh, and R.E. Welsch, *Regression Diagnostics*, John Wiley, New York, 1980.
- [4] K.N. Berk, *Tolerance and condition in regression computations*, J. Am. Stat. Assoc. 72 (1977), pp. 863–866.
- [5] J. Dias and J. Castro, *New multicollinearity indicators in linear regression models*, Int. Stat. Rev. 75 (2007), pp. 114–121.

- [6] Economic Reports of the President, 2011. Available at <http://www.gpo.gov/fdsys/browse>
- [7] J. Fox, *Applied Regression Analysis, Linear Models, and Related Methods*, Sage, Thousand Oaks, CA, 1997.
- [8] J. Fox and G. Monette, *Generalized collinearity diagnostics*, J. Am. Stat. Assoc. 87 (1992), pp. 178–183.
- [9] M. Friendly, The generalized ridge trace plot: Visualizing bias and precision. J. Comput. Graph. Statist. 22 (2013), pp. 50–68.
- [10] M. Friendly and E. Kwan, *Where's Waldo? Visualizing collinearity diagnostics*, Am. Stat. 63 (2009), pp. 56–65.
- [11] J. García, C. García, M.D.M. López, and R. Salmerón, *Desarrollo del método de alzado para el tratamiento de la multicolinealidad, Determinación del factor de alzado*, Anales de Economía Aplicada, Zaragoza, 2013.
- [12] C. García, J. García, M.D.M. López, and R. Salmerón, *Collinearity: Revisiting the variance inflation factor in ridge regression*, J. Appl. Stat. 32 (2015), pp. 648–661. doi:10.1080/02664763.2014.980789
- [13] J. García, R. Salmerón, C. García, and M.D.M. López, *Standardization of variables and collinearity diagnostic in ridge regression*, Int. Stat. Rev., in press. doi:10.1111/insr.12099
- [14] J. García, R. Salmerón, M.D.M. López, and C. García, *Revisiting the Condition Number and Red Indicator in Ridge Regression*, Proceedings of the International Work Conference on Time Series, Granada, 2015, pp. 271–22.
- [15] D.G. Gibbons, *A simulation study of some ridge estimators*, J. Am. Stat. Assoc. 76 (1981), pp. 131–139.
- [16] A.E. Hoerl and R.W. Kennard, *Ridge regression: Applications to nonorthogonal problems*, Technometrics 12 (1970), pp. 69–82.
- [17] A.E. Hoerl and R.W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics 12 (1970), pp. 55–67.
- [18] D.R. Jensen and D.E. Ramirez, *Surrogate models in ill-conditioned systems*, J. Stat. Plan. Inference 140 (2010), pp. 2069–2077.
- [19] P.A. Kennedy, *A Guide to Econometrics*, 4th ed. MIT Press, Cambridge, MA, 1992.
- [20] P. Kovacs, T. Petres, and L. Toth, *A new measure of multicollinearity in linear regression models*, Int. Stat. Rev. 73 (2005), pp. 405–412.
- [21] M.H. Kutner, C.J. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*, 5th ed. McGraw-Hill, New York, 2005.
- [22] K. Levenberg, *A method for the solution of certain non-linear problems in least squares*, Quart. Appl. Math. 2 (1944), pp. 164–168.
- [23] D.W. Marquardt, *Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation*, Technometrics 12 (1970), pp. 591–612.
- [24] G.C. McDonald, *Ridge regression*, Wiley Interdiscip. Rev. Comput. Stat. 1 (2009), pp. 93–100.
- [25] G.C. McDonald, *Tracing ridge regression coefficients*, Wiley Interdiscip. Rev.: Comput. Stat. 2 (2010), pp. 695–703.
- [26] G.C. McDonald and D.I. Galarneau, *A Monte Carlo evaluation of some ridge type estimators*, J. Am. Stat. Assoc. 70 (1975), pp. 407–416.
- [27] R.M. O'Brien, *A caution regarding rules of thumb for variance inflation factors*, Qual. Quant. 41 (2007), pp. 673–690.
- [28] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, 2012. ISBN:3-900051-07-0
- [29] J.O. Rawlings, S.G. Pantula, and D.A. Dickey, *Applied Regression Analysis: A Research Tool*, Springer Science & Business Media, New York, 1998.
- [30] J.D. Riley, *Solving systems of linear equations with a positive definite, symmetric but possibly ill-conditioned matrix*, Math. Tables Aids Comput. 9 (1955), pp. 96–101.

- [31] R. Salmerón, C. García, M.D.M. López, and J. García, *A Note About the Variance Inflation Factor and the Ridge Regression*, 2nd International Conference of Informatics and Management Sciences, Slovak Republic, 2013, pp. 197–199.
- [32] H. Theil, *Principles of Econometrics*, Wiley, New York, 1971.
- [33] D.W. Wichern and G.A. Churchill, *A comparison of ridge estimators*, *Technometrics* 20 (1978), pp. 301–311.
- [34] C.R. Wichers, *The detection of multicollinearity: A comment*, *Rev. Econ. Stat.* 57 (1975), pp. 366–368.
- [35] J. Wissel, *A new biased estimator for multivariate regression models with highly collinear variables*, Ph.D. thesis, Dissertation zur Erlangung des naturwissenschaftlichen Doktorgrades der Bayerischen Julius-Maximilians-Universität Würzburg, 2009.

Appendix 1. Relationship between the definition of the VIF by Kutner *et al.* [21] and the extension of the Marquardt's VIF definition in RE

Kutner *et al.* [21] considered that in RE the VIF is the i th element of the principal diagonal of the following matrix:

$$\mathbf{V} \begin{pmatrix} \frac{\gamma_1}{(\gamma_1 + \lambda)^2} & 0 & \dots & 0 \\ 0 & \frac{\gamma_2}{(\gamma_2 + \lambda)^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\gamma_p}{(\gamma_p + \lambda)^2} \end{pmatrix} \mathbf{V}', \quad (\text{A1})$$

where \mathbf{V} is an orthogonal matrix of order p whose columns are the normalized eigenvectors of the correlation matrix of the independent variables, \mathbf{R} , and γ_i is the i th eigenvalue of the same matrix, $i = 1, \dots, p$. That is to say, $\mathbf{R} = \mathbf{V} \cdot \mathbf{\Gamma} \cdot \mathbf{V}'$ with $\mathbf{V}' \cdot \mathbf{V} = \mathbf{V} \cdot \mathbf{V}' = \mathbf{I}$ and $\mathbf{\Gamma}$ being a diagonal matrix whose elements are the eigenvalues of \mathbf{R} . Since the data are standardized, $\mathbf{R} = \mathbf{X}'\mathbf{X}$, and then this proposal coincides with the one provided by Marquardt since

$$(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1} = \mathbf{V}(\mathbf{\Gamma} + \lambda\mathbf{I})^{-1}\mathbf{\Gamma}(\mathbf{\Gamma} + \lambda\mathbf{I})^{-1}\mathbf{V}', \quad (\text{A2})$$

where

$$(\mathbf{\Gamma} + \lambda\mathbf{I})^{-1}\mathbf{\Gamma}(\mathbf{\Gamma} + \lambda\mathbf{I})^{-1} = \begin{pmatrix} \frac{\gamma_1}{(\gamma_1 + \lambda)^2} & 0 & \dots & 0 \\ 0 & \frac{\gamma_2}{(\gamma_2 + \lambda)^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{\gamma_p}{(\gamma_p + \lambda)^2} \end{pmatrix}. \quad (\text{A3})$$

Appendix 2. VIFs in RE

In this section we obtain explicit expression for RESVIF, REVIF and REGVIF for $p = 3$.

A.1 Ridge estimator surrogate VIF (RESVIF)

From surrogate ridge estimator presented by Jensen and Ramirez [18], we can affirm that there exists a matrix \mathbf{X}_S' which verifies $\mathbf{X}_S'\mathbf{X}_S = \mathbf{X}'\mathbf{X} + k\mathbf{I}$. In the case of three variables ($p = 3$)

$$\mathbf{X}_S'\mathbf{X}_S = \mathbf{X}'\mathbf{X} + \lambda\mathbf{I} = \begin{pmatrix} 1 + \lambda & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 + \lambda & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 + \lambda \end{pmatrix}, \quad (\text{A4})$$

i.e. $\sum_{i=1}^n x_{ij}^2 = 1 + \lambda$ for $j = 1, 2, 3$, and $\sum_{i=1}^n x_{ij}x_{ik} = \rho_{jk}$ for $j, k = 1, 2, 3, j \neq k$.

Through the regression of the standardized variable $\mathbf{X}_S(i)$ on the standardized variables $\mathbf{X}_S(j)$ and $\mathbf{X}_S(k)$

$$\mathbf{X}_S(i) = \beta_j \mathbf{X}_S(j) + \beta_k \mathbf{X}_S(k) + \mathbf{v}, \quad (\text{A5})$$

with $i, j, k \in \{1, 2, 3\}$, $i \neq j$, $i \neq k$ and $j \neq k$, the estimator of the parameters by OLS is

$$\hat{\beta}_S(i) = \begin{pmatrix} 1 + \lambda & \rho_{jk} \\ \rho_{jk} & 1 + \lambda \end{pmatrix}^{-1} \begin{pmatrix} \rho_{ij} \\ \rho_{ik} \end{pmatrix} = \frac{1}{(1 + \lambda)^2 - \rho_{jk}^2} \begin{pmatrix} (1 + \lambda)\rho_{ij} - \rho_{jk}\rho_{ik} \\ (1 + \lambda)\rho_{ik} - \rho_{jk}\rho_{ij} \end{pmatrix}. \quad (\text{A6})$$

In this case the explained sum of squares (ESS) and the total sum of squared (TSS) are

$$\text{ESS}_S(i) = \frac{1}{(1 + \lambda)^2 - \rho_{jk}^2} [(1 + \lambda)(\rho_{ij}^2 + \rho_{ik}^2) - 2\rho_{ij}\rho_{jk}\rho_{ik}], \quad (\text{A7})$$

$$\text{TSS}_S(i) = 1 + \lambda \quad (\text{by } \mathbf{X}_S \text{ definition}). \quad (\text{A8})$$

Using the expressions (A7) and (A8), the determination coefficient of the model (A5) is

$$R_S^2(i) = \frac{\text{ESS}_S(i)}{\text{TSS}_S(i)} = \frac{(1 + \lambda)(\rho_{ij}^2 + \rho_{ik}^2) - 2\rho_{ij}\rho_{jk}\rho_{ik}}{(1 + \lambda)[(1 + \lambda)^2 - \rho_{jk}^2]}. \quad (\text{A9})$$

As a result, the i th variable RESVIF is

$$\text{RESVIF}(\lambda, i) = \frac{1}{1 - R_S^2(i)} = \frac{(1 + \lambda)[(1 + \lambda)^2 - \rho_{jk}^2]}{(1 + \lambda)[(1 + \lambda)^2 - \rho_{ij}^2 - \rho_{ik}^2 - \rho_{jk}^2] + 2\rho_{ij}\rho_{jk}\rho_{ik}}. \quad (\text{A10})$$

Because of $\text{ESS}_S(i) \geq 0$ and $(1 + \lambda)^2 - \rho_{jk}^2 \geq 0$ then $2\rho_{ij}\rho_{jk}\rho_{ik} - (1 + \lambda)(\rho_{ij}^2 + \rho_{ik}^2) \leq 0$. By adding $(1 + \lambda)[(1 + \lambda)^2 - \rho_{jk}^2]$ on both sides we obtain

$$(1 + \lambda)[(1 + \lambda)^2 - \rho_{ij}^2 - \rho_{ik}^2 - \rho_{jk}^2] + 2\rho_{ij}\rho_{jk}\rho_{ik} \leq (1 + \lambda)[(1 + \lambda)^2 - \rho_{jk}^2].$$

Thus, we conclude that $\text{RESVIF}(\lambda, i) \geq 1$, $\lim_{\lambda \rightarrow \infty} \text{RESVIF}(\lambda, i) = 1$ and the RESVIF is monotone decreasing with increasing λ (see [18, Theorem 5, p. 2077]).

A.2 Ridge estimator VIF (REVIF)

Ridge estimator VIF (REVIF) is the extension of Theil [32] definition (see expression (2)) to RE. In this case, it is necessary to know the matrix of regressors \mathbf{Z} so that $\mathbf{Z}'\mathbf{Z} = \mathbf{X}'\mathbf{X} + \lambda\mathbf{I}$. Even though this matrix is unknown Marquardt [23] showed that

$$\mathbf{X}_Z = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \\ \hline \sqrt{\lambda} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I} \end{pmatrix}, \quad \mathbf{Y}_Z = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} \mathbf{Y} \\ \mathbf{0} \end{pmatrix},$$

where the bottom $p \times 1$ submatrix of \mathbf{Y}_Z is not to be viewed as a random responses, the top $n \times p$ submatrix of the \mathbf{X}_Z matrix has to be standardized (see [13]) and RE in the model (1) is similar to the OLS regression in the following model:

$$\mathbf{Y}_Z = \mathbf{X}_Z\boldsymbol{\beta} + \mathbf{v}. \quad (\text{A11})$$

It is verifying that $\hat{\boldsymbol{\beta}}(\lambda) = (\mathbf{X}_Z'\mathbf{X}_Z)^{-1}\mathbf{X}_Z'\mathbf{Y}_Z = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$ since $\mathbf{X}_Z'\mathbf{X}_Z = \mathbf{X}'\mathbf{X} + \lambda\mathbf{I}$ and $\mathbf{X}_Z'\mathbf{Y}_Z = \mathbf{X}'\mathbf{Y}$. Then the matrix \mathbf{X}_Z can be used instead of the matrix \mathbf{Z} . The case $p = 2$ is given in

García *et al.* [12]. For $p = 3$, the regression of the variable $\mathbf{X}_Z(i)$ on the variables $\mathbf{X}_Z(j)$ and $\mathbf{X}_Z(k)$ are given by

$$\mathbf{X}_Z(i) = \beta_0 + \beta_j \mathbf{X}_Z(j) + \beta_k \mathbf{X}_Z(k) + \mathbf{w}, \quad (\text{A12})$$

with $i, j, k \in \{1, 2, 3\}$, $i \neq j$, $i \neq k$ and $j \neq k$. In that case

$$\text{ESS}_Z(i) = \frac{\lambda[(\lambda + \rho_{jk} + 1) - 2\rho_{ij} - 2\rho_{ik}]}{B} - \frac{2\rho_{ij}\rho_{ik}D - (\rho_{ij}^2 + \rho_{ik}^2)C}{A}, \quad (\text{A13})$$

$$\text{TSS}_Z(i) = \frac{n + 3 + \lambda(n + 2)}{n + 3}, \quad (\text{A14})$$

where

$$A = \lambda^2(n + 1) + \lambda(4 + 2n + 2\rho_{jk}) + (n + 3)(1 - \rho_{jk}^2),$$

$$B = (n + 3)(1 + \rho_{jk}) + \lambda(n + 1),$$

$$C = n + 3 + \lambda(n + 2),$$

$$D = \rho_{jk}(n + 3) - \lambda.$$

With this information, we conclude that the coefficient of determination is

$$R_Z^2(i) = \frac{\text{ESS}_Z(i)}{\text{TSS}_Z(i)} = \frac{n + 3}{n + 3 + \lambda(n + 2)} \text{ESS}_Z(i), \quad (\text{A15})$$

and then the i th variable REVIF is

$$\text{REVIF}(\lambda, i) = \frac{n + 3 + \lambda(n + 2)}{(n + 3) \left[\frac{\lambda(2\rho_{ij} + 2\rho_{ik} - (\lambda + \rho_{jk} - n)) + (n + 3)(\rho_{jk} + 1)}{B} + \frac{2\rho_{ij}\rho_{ik}D - (\rho_{ij}^2 + \rho_{ik}^2)C}{A} \right] + \lambda(n + 2)}. \quad (\text{A16})$$

It can be easily demonstrated for $\lambda = 0$ that

$$\text{RESVIF}(0, i) = \text{REVIF}(0, i) = \frac{1 - \rho_{jk}^2}{1 - \rho_{jk}^2 - \rho_{ij}^2 - \rho_{ik}^2 + 2\rho_{jk}\rho_{ij}\rho_{ik}} = \text{VIF}(i). \quad (\text{A17})$$

Thus, we conclude that $\text{REVIF}(\lambda, i) \geq 1$, $\lim_{\lambda \rightarrow \infty} \text{REVIF}(\lambda, i) = 1$ and the REVIF is monotone decreasing with increasing λ since $R_Z^2(i)$ is decreasing in λ (see [25, p. 696]).

On the other hand, with the help of symbolic computation software we can affirm that

$$\lim_{n \rightarrow \infty} (\text{REVIF}(\lambda, i) - \text{RESVIF}(\lambda, i)) = 0. \quad (\text{A18})$$

A.3 Generalized VIF for $p = 3$ (REGVIF)

When $p = 3$ the correlation matrix among all variables from Equation (6) is

$$\mathbf{R} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}. \quad (\text{A19})$$

In this case the expression (7) would be written as

$$\text{REGVIF}(i) = \frac{1 - \rho_{jk}^2}{1 - \rho_{jk}^2 - \rho_{ij}^2 - \rho_{ik}^2 + 2\rho_{jk}\rho_{ij}\rho_{ik}}, \quad (\text{A20})$$

with $i, j, k = 1, 2, 3$, $i \neq j$, $i \neq k$, $j \neq k$, since $|\mathbf{R}| = 1 - \rho_{jk}^2 - \rho_{ij}^2 - \rho_{ik}^2 + 2\rho_{jk}\rho_{ij}\rho_{ik}$, $\mathbf{R}_{-i} = \begin{pmatrix} 1 & \rho_{jk} \\ \rho_{jk} & 1 \end{pmatrix}$ and $|\mathbf{R}_{-i}| = 1 - \rho_{jk}^2$.

Therefore, the expression (34) clearly verifies that $\text{RESVIF}(0, i) = \text{REVIF}(0, i) = \text{REGVIF}(i)$ (see expression (31)). Also $\text{REGVIF}(i) \geq 1, \forall i$, since $\text{REGVIF}(i) = \text{RESVIF}(0, i) \geq 1$. Contrarily to RESVIF and REVIF , the REGVIF will not be decreasing in λ .

For $\lambda = 0$ it is clear what $\mathbf{R}(0) = \mathbf{R}$ is, but the question is how to calculate $\mathbf{R}(\lambda)$ for $\lambda > 0$. Friendly [9] solves this problem by transforming the covariance matrix (6) of RE into a correlation matrix considering this last one as $\mathbf{R}(\lambda)$.

Note that if we consider the following matrix:

$$\mathbf{R}(\lambda) = \begin{pmatrix} 1 + \lambda & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 + \lambda & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 + \lambda \end{pmatrix}, \quad (\text{A21})$$

and applying the definition (4) we obtain that

$$\text{REGVIF}(\lambda, i) = \frac{(1 + \lambda)[(1 + \lambda)^2 - \rho_{jk}^2]}{(1 + \lambda)[(1 + \lambda)^2 - \rho_{ij}^2 - \rho_{ik}^2 - \rho_{jk}^2] + 2\rho_{ij}\rho_{jk}\rho_{ik}}.$$

It is to say, $\text{REGVIF}(\lambda, i) = \text{RESVIF}(\lambda, i)$ for all λ and i . However, in this case the election of $\mathbf{R}(\lambda)$ will not correspond to a true correlation matrix since its mean diagonal is not constantly 1 for $\lambda > 0$.

Appendix 3. Algorithms to obtain the VIFs in RE in the general case

Consider that we have p standardized exogenous variables and an endogenous variable, if we consider the multiple linear regression model (1), the surrogate and ridge estimator VIFs for any value of p may be computed as shown in the Algorithms 1 and 2. In both cases, the matrices \mathbf{X}_S and \mathbf{X}_Z are calculated to obtain the regressions (19) and (27), respectively.

Finally, the coefficients of determination of these regressions are used to obtain the VIFs. Note that these algorithms simply reproduce the steps followed in Sections 6 and 6. The generalized VIF is obtained from expression (7) as shown in Algorithm 3. We have also implemented an iterative procedure to calculate the VIF given by Marquardt [23], although it is not the aim of the work. In Algorithm 4 the main diagonal elements of matrix $(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}$ are obtained. Upon request, generate codes are available from the authors.

Algorithm 1 Obtaining surrogate VIF (RESVIF)

Require: Calculate $\mathbf{X}'\mathbf{X}$, \mathbf{I} and $D(\delta)$ (discretization of the interval $[0, 1]$ with δ points)

- 1: **for** $\lambda \in D(\delta)$ **do**
 - 2: consider the surrogate matrix \mathbf{X}_S so that $\mathbf{X}_S'\mathbf{X}_S = \mathbf{X}'\mathbf{X} + \lambda\mathbf{I}$
 - 3: **for** $i \in \{1, 2, \dots, p\}$ **do**
 - 4: the regression of column i of \mathbf{X}_S on the other variables
 - 5: obtain the coefficient of determination, $R_S^2(i)$, of the regression
 - 6: calculate surrogate VIF of i th variable as $\text{RESVIF}(\lambda, i) = \frac{1}{1 - R_S^2(i)}$
 - 7: **end for**
 - 8: **end for**
-

Algorithm 2 Obtaining ridge estimator VIF (REVIF)

Require: Calculate $\mathbf{X}'\mathbf{X}$, \mathbf{I} and $D(\delta)$ (discretization of the interval $[0, 1]$ with δ points)

- 1: **for** $\lambda \in D(\delta)$ **do**
 - 2: calculate the matrix \mathbf{X}_Z that generates the matrix of the ridge estimators $\mathbf{X}'_Z\mathbf{X}_Z = \mathbf{X}'\mathbf{X} + \lambda\mathbf{I}$
 - 3: **for** $i \in \{1, 2, \dots, p\}$ **do**
 - 4: the regression of column i of \mathbf{X}_Z on the other variables
 - 5: obtain the coefficient of determination, $R_Z^2(i)$, of the regression
 - 6: calculate surrogate VIF of i th variable as $\text{REVIF}(\lambda, i) = \frac{1}{1-R_Z^2(i)}$
 - 7: **end for**
 - 8: **end for**
-

Algorithm 3 Obtaining generalized VIF (REGVIF)

Require: Calculate $\mathbf{X}'\mathbf{X}$, \mathbf{I} , $D(\delta)$ (discretization of the interval $[0, 1]$ with δ points) and $\hat{\sigma}^2$ (the estimation of the variance of regression)

- 1: **for** $\lambda \in D(\delta)$ **do**
 - 2: calculate $\text{var}(\hat{\boldsymbol{\beta}}(\lambda)) = \hat{\sigma}^2 (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1} \mathbf{X}'\mathbf{X} (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}$
 - 3: transform $\text{var}(\hat{\boldsymbol{\beta}}(\lambda))$ into a correlation matrix \mathbf{R}
 - 4: **for** $i \in \{1, 2, \dots, p\}$ **do**
 - 5: calculate generalized VIF of i th variable as $\text{REGVIF}(i) = \frac{|\mathbf{R}_{-i}|}{|\mathbf{R}|}$
 - 6: **end for**
 - 7: **end for**
-

Algorithm 4 Obtaining Marquardt VIF (REMVIF)

Require: Calculate $\mathbf{X}'\mathbf{X}$, \mathbf{I} and $D(\delta)$ (discretization of the interval $[0, 1]$ with δ points)

- 1: **for** $\lambda \in D(\delta)$ **do**
 - 2: calculate $\mathbf{M}(\lambda) = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1} \mathbf{X}'\mathbf{X} (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}$
 - 3: **end for**
 - 4: consider the diagonal elements of $\mathbf{M}(\lambda)$ as REMVIF
-

Appendix 4. Data empirical application

Table A1. Data for empirical application.

| Year | Y | X_1 | X_2 | X_3 |
|------|----------|---------|----------|---------|
| 1995 | 4524.80 | 6076.20 | 6200.90 | 1140.74 |
| 1996 | 4792.40 | 6288.30 | 6591.60 | 1253.43 |
| 1997 | 5104.80 | 6520.40 | 7000.70 | 1324.76 |
| 1998 | 5589.50 | 6862.30 | 7525.40 | 1420.99 |
| 1999 | 6195.10 | 7237.60 | 7910.80 | 1531.10 |
| 2000 | 6752.60 | 7604.60 | 8559.40 | 1716.96 |
| 2001 | 7460.40 | 7810.30 | 8883.30 | 1867.85 |
| 2002 | 8361.20 | 8018.30 | 9060.10 | 1972.11 |
| 2003 | 9376.20 | 8244.50 | 9378.10 | 2077.36 |
| 2004 | 10650.70 | 8515.80 | 9937.20 | 2192.24 |
| 2005 | 12097.70 | 8803.50 | 10485.90 | 2290.93 |
| 2006 | 13481.90 | 9054.50 | 11268.10 | 2384.96 |
| 2007 | 14566.00 | 9262.90 | 11912.30 | 2528.77 |
| 2008 | 14661.30 | 9211.70 | 12460.20 | 2548.86 |
| 2009 | 14370.00 | 9032.60 | 11867.00 | 2438.73 |
| 2010 | 13712.30 | 9196.20 | 12321.90 | 2545.28 |
| 2011 | 13383.80 | 9428.80 | 12947.30 | 2631.51 |

Source: Economic Reports of the President [6].

Copyright of Journal of Applied Statistics is the property of Routledge and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.